

Factor Analysis

Richard B. Darlington

Factor analysis includes both *component analysis* and *common factor analysis*. More than other statistical techniques, factor analysis has suffered from confusion concerning its very purpose. This affects my presentation in two ways. First, I devote a long section to describing what factor analysis does before examining in later sections how it does it. Second, I have decided to reverse the usual order of presentation. Component analysis is simpler, and most discussions present it first. However, I believe common factor analysis comes closer to solving the problems most researchers actually want to solve. Thus learning component analysis first may actually interfere with understanding what those problems are. Therefore component analysis is introduced only quite late in this chapter.

What Factor Analysis Can and Can't Do

I assume you have scores on a number of variables-- anywhere from 3 to several hundred variables, but most often between 10 and 100. Actually you need only the correlation or covariance matrix--not the actual scores. The purpose of factor analysis is to discover simple patterns in the pattern of relationships among the variables. In particular, it seeks to discover if the observed variables can be explained largely or entirely in terms of a much smaller number of variables called *factors*.

Some Examples of Factor-Analysis Problems

1. Factor analysis was invented nearly 100 years ago by psychologist Charles Spearman, who hypothesized that the enormous variety of tests of mental ability--measures of mathematical skill, vocabulary, other verbal skills, artistic skills, logical reasoning ability, etc.--could all be explained by one underlying "factor" of general intelligence that he called *g*. He hypothesized that if *g* could be measured and you could select a subpopulation of people with the same score on *g*, in that subpopulation you would find no correlations among any tests of mental ability. In other words, he hypothesized that *g* was the only factor common to all those measures.

It was an interesting idea, but it turned out to be wrong. Today the College Board testing service operates a system based on the idea that there are at least three important factors of mental ability--verbal, mathematical, and logical abilities--and most psychologists agree that many other factors could be identified as well.

2. Consider various measures of the activity of the autonomic nervous system--heart rate, blood pressure, etc. Psychologists have wanted to know whether, except for random fluctuation, all those measures move up and down together--the "activation" hypothesis. Or do groups of autonomic measures move up and down together, but separate from other groups? Or are all the measures largely independent? An unpublished analysis of mine found that in one data set, at any rate, the data fitted the activation hypothesis quite well.

3. Suppose many species of animal (rats, mice, birds, frogs, etc.) are trained that food will appear at a certain spot whenever a noise--any kind of noise--comes from that spot. You could then tell whether they could detect a particular sound by seeing whether they turn in that direction when the sound appears. Then if you studied many sounds and many species, you might want to know on how many different dimensions of hearing acuity the species vary. One hypothesis would be that they vary on just three dimensions--the ability to detect high-frequency sounds, ability to detect low-frequency sounds,

and ability to detect intermediate sounds. On the other hand, species might differ in their auditory capabilities on more than just these three dimensions. For instance, some species might be better at detecting sharp click-like sounds while others are better at detecting continuous hiss-like sounds.

4. Suppose each of 500 people, who are all familiar with different kinds of automobiles, rates each of 20 automobile models on the question, "How much would you like to own that kind of automobile?" We could usefully ask about the number of dimensions on which the ratings differ. A one-factor theory would posit that people simply give the highest ratings to the most expensive models. A two-factor theory would posit that some people are most attracted to sporty models while others are most attracted to luxurious models. Three-factor and four-factor theories might add safety and reliability. Or instead of automobiles you might choose to study attitudes concerning foods, political policies, political candidates, or many other kinds of objects.

5. Rubenstein (1986) studied the nature of curiosity by analyzing the agreements of junior-high-school students with a large battery of statements such as "I like to figure out how machinery works" or "I like to try new kinds of food." A factor analysis identified seven factors: three measuring enjoyment of problem-solving, learning, and reading; three measuring interests in natural sciences, art and music, and new experiences in general; and one indicating a relatively low interest in money.

The Goal: Understanding of Causes

Many statistical methods are used to study the relation between independent and dependent variables. Factor analysis is different; it is used to study the patterns of relationship among many dependent variables, with the goal of discovering something about the nature of the independent variables that affect them, even though those independent variables were not measured directly. Thus answers obtained by factor analysis are necessarily more hypothetical and tentative than is true when independent variables are observed directly. The inferred independent variables are called *factors*. A typical factor analysis suggests answers to four major questions:

1. How many different factors are needed to explain the pattern of relationships among these variables?
2. What is the nature of those factors?
3. How well do the hypothesized factors explain the observed data?
4. How much purely random or unique variance does each observed variable include?

I illustrate these questions later.

Absolute Versus Heuristic Uses of Factor Analysis

A *heuristic* is a way of thinking about a topic which is convenient even if not absolutely true. We use a heuristic when we talk about the sun rising and setting as if the sun moved around the earth, even though we know it doesn't. "Heuristic" is both a noun and an adjective; to use a heuristic is to think in heuristic terms.

The previous examples can be used to illustrate a useful distinction--between *absolute* and *heuristic* uses of factor analysis. Spearman's *g* theory of intelligence, and the activation theory of autonomic functioning, can be thought of as absolute theories which are or were hypothesized to give complete descriptions of the pattern of relationships among variables. On the other hand, Rubenstein never claimed that her list of the seven major factors of curiosity offered a complete description of curiosity. Rather those factors merely appear to be the most important seven factors--the best way of summarizing a body of data. Factor analysis can suggest either absolute or heuristic models; the distinction is in how you interpret the output.

Is Factor Analysis Objective?

The concept of heuristics is useful in understanding a property of factor analysis which confuses many people. Several scientists may apply factor analysis to similar or even identical sets of measures, and one may come up with 3 factors while another comes up with 6 and another comes up with 10. This lack of agreement has tended to discredit all uses of factor analysis. But if three travel writers wrote travel guides to the United States, and one divided the country into 3 regions, another into 6, and another into 10, would we say that they contradicted each other? Of course not; the various writers are just using convenient ways of organizing a topic, not claiming to represent the only correct way of doing so. Factor analysts reaching different conclusions contradict each other only if they all claim absolute theories, not heuristics. The fewer factors the simpler the theory; the more factors the better the theory fits the data. Different workers may make different choices in balancing simplicity against fit.

A similar balancing problem arises in regression and analysis of variance, but it generally doesn't prevent different workers from reaching nearly or exactly the same conclusions. After all, if two workers apply an analysis of variance to the same data, and both workers drop out the terms not significant at the .05 level, then both will report exactly the same effects. However, the situation in factor analysis is very different. For reasons explained later, there is no significance test in component analysis that will test a hypothesis about the number of factors, as that hypothesis is ordinarily understood. In common factor analysis there is such a test, but its usefulness is limited by the fact that it frequently yields more factors than can be satisfactorily interpreted. Thus a worker who wants to report only interpretable factors is still left without an objective test.

A similar issue arises in identifying the nature of the factors. Two workers may each identify 6 factors, but the two sets of factors may differ--perhaps substantially. The travel-writer analogy is useful here too; two writers might each divide the US into 6 regions, but define the regions very differently.

Another geographical analogy may be more parallel to factor analysis, since it involves computer programs designed to maximize some quantifiable objective. Computer programs are sometimes used to divide a state into congressional districts which are geographically contiguous, nearly equal in population, and perhaps homogeneous on dimensions of ethnicity or other factors. Two different district-creating programs might come up with very different answers, though both answers are reasonable. This analogy is in a sense too good; we believe that factor analysis programs usually don't yield answers as different from each other as district-creating programs do.

Factor Analysis Versus Clustering and Multidimensional Scaling

Another challenge to factor analysis has come from the use of competing techniques such as cluster analysis and multidimensional scaling. While factor analysis is typically applied to a correlation matrix, those other methods can be applied to any sort of matrix of similarity measures, such as ratings of the similarity of faces. But unlike factor analysis, those methods cannot cope with certain unique properties of correlation matrices, such as reflections of variables. For instance, if you reflect or reverse the scoring direction of a measure of "introversion", so that high scores indicate "extroversion" instead of introversion, then you reverse the signs of all that variable's correlations: $-.36$ becomes $+.36$, $+.42$ becomes $-.42$, and so on. Such reflections would completely change the output of a cluster analysis or multidimensional scaling, while factor analysis would recognize the reflections for what they are; the reflections would change the signs of the "factor loadings" of any reflected variables, but would not change anything else in the factor analysis output.

Another advantage of factor analysis over these other methods is that factor analysis can recognize certain properties of correlations. For instance, if variables A and B each correlate .7 with variable C, and correlate .49 with each other, factor analysis can recognize that A and B correlate zero when C is held constant because $.7^2 = .49$. Multidimensional scaling and cluster analysis have no ability to recognize such relationships, since the correlations are treated merely as generic "similarity measures" rather than as correlations.

We are not saying these other methods should never be applied to correlation matrices; sometimes they yield insights not available through factor analysis. But they have definitely not made factor analysis obsolete. The next section touches on this point.

Factors "Differentiating" Variables Versus Factors "Underlying" Variables

When someone says casually that a set of variables seems to reflect "just one factor", there are several things they might mean that have nothing to do with factor analysis. If we word statements more carefully, it turns out that the phrase "just one factor *differentiates* these variables" can mean several different things, none of which corresponds to the factor analytic conclusion that "just one factor *underlies* these variables".

One possible meaning of the phrase about "differentiating" is that a set of variables all correlate highly with each other but differ in their means. A rather similar meaning can arise in a different case. Consider several tests A, B, C, D which test the same broadly-conceived mental ability, but which increase in difficulty in the order listed. Then the highest correlations among the tests may be between adjacent items in this list (r_{AB} , r_{BC} and r_{CD}) while the lowest correlation is between items at the opposite ends of the list (r_{AD}). Someone who observed this pattern in the correlations among the items might well say the tests "can be put in a simple order" or "differ in just one factor", but that conclusion has nothing to do with factor analysis. This set of tests would *not* contain just one common factor.

A third case of this sort may arise if variable A affects B, which affects C, which affects D, and those are the only effects linking these variables. Once again, the highest correlations would be r_{AB} , r_{BC} and r_{CD} while the lowest correlation would be r_{AD} . Someone might use the same phrases just quoted to describe this pattern of correlations; again it has nothing to do with factor analysis.

A fourth case is in a way a special case of all the previous cases: a perfect Guttman scale. A set of dichotomous items fits a Guttman scale if the items can be arranged so that a negative response to any item implies a negative response to all subsequent items while a positive response to any item implies a positive response to all previous items. For a trivial example consider the items

- Are you above 5 feet 2 inches in height?
- Are you above 5 feet 4 inches in height?
- Are you above 5 feet 6 inches in height?
- Etc.

To be consistent, a person answering negatively to any of these items must answer negatively to all later items, and a positive answer implies that all previous answers must be positive. For a nontrivial example consider the following questionnaire items:

- Should our nation lower tariff barriers with nation B?
- Should our two central banks issue a single currency?
- Should our armies become one?
- Should we fuse with nation B, becoming one nation?

If it turned out that these items formed a perfect Guttman scale, it would be easier to describe peoples' attitudes about "nation B" than if they didn't. When a set of items does form a Guttman scale, interestingly it does not imply that factor analysis would discover a single common factor. A Guttman scale implies that one factor *differentiates* a set of items (e.g, "favorableness toward cooperation with nation B"), not that one factor *underlies* those items.

Applying multidimensional scaling to a correlation matrix could discover all these simple patterns of differences among variables. Thus multidimensional scaling seeks factors which *differentiate* variables while factor analysis looks for the factors which *underlie* the variables. Scaling may sometimes find simplicity where factor analysis finds none, and factor analysis may find simplicity where scaling finds none.

A Dubious History

If a statistical method can have an embarrassing history, factor analysis is that method. Around 1950 the reputation of factor analysis suffered from overpromotion by a few overenthusiastic partisans. In retrospect there were three things wrong with the way some people were thinking about factor analysis at that time. First, some people seemed to see factor analysis as *the* statistical method rather than *a* statistical method. Second, they were thinking in absolute terms about problems for which a heuristic approach would have been more appropriate. Third, they were thinking of overly broad sets of variables ("we want to understand all of human personality" rather than "we want to understand the nature of curiosity"). Thus in three different ways, they were attempting to stretch factor analysis farther than it was capable of going. In recent decades factor analysis seems to have found its rightful place as a family of methods which is useful for certain limited purposes.

Basic Concepts and Principles

A Simple Example

A factor analysis usually begins with a correlation matrix I'll denote R. Below is an artificial 5 x 5 correlation matrix I'll call R55.

1.00	.72	.63	.54	.45
.72	1.00	.56	.48	.40
.63	.56	1.00	.42	.35
.54	.48	.42	1.00	.30
.45	.40	.35	.30	1.00

Imagine that these are correlations among 5 variables measuring mental ability. Matrix R55 is exactly consistent with the hypothesis of a single common factor *g* whose correlations with the 5 observed variables are respectively .9, .8, .7, .6, and .5. To see why, consider the formula for the partial correlation between two variables *a* and *b* partialing out a third variable *g*:

$$r_{ab.g} = (r_{ab} - r_{ag} r_{bg}) / \sqrt{[(1-r_{ag}^2)(1-r_{bg}^2)]}$$

This formula shows that $r_{ab.g} = 0$ if and only if $r_{ab} = r_{ag} r_{bg}$. The requisite property for a variable to function as a general factor *g* is that any partial correlation between any two observed variables, partialing out *g*, is zero. Therefore if a correlation matrix can be explained by a general factor *g*, it will be true that there is some set of correlations of the observed variables with *g*, such that the product of any two of those correlations equals the correlation between the two observed variables. But matrix

R55 has exactly that property. That is, any off-diagonal entry r_{jk} is the product of the j th and k th entries in the row .9 .8 .7 .6 .5. For instance, the entry in row 1 and column 3 is $.9 \times .7$ or $.63$. Thus matrix R55 exactly fits the hypothesis of a single common factor.

If we found that pattern in a real correlation matrix, what exactly would we have shown? First, the existence of the factor is *inferred* rather than *observed*. We certainly wouldn't have *proven* that scores on these 5 variables are affected by just one common factor. However, that is the simplest or most parsimonious hypothesis that fits the pattern of observed correlations.

Second, we would have an estimate of the factor's correlation with each of the observed variables, so we can say something about the factor's nature, at least in the sense of what it correlates highly with or doesn't correlate with. In this example the values .9 .8 .7 .6 .5 are these estimated correlations.

Third, we couldn't measure the factor in the sense of deriving each person's exact score on the factor. But we can if we wish use methods of multiple regression to estimate each person's score on the factor from their scores on the observed variables.

Matrix R55 is virtually the simplest possible example of common factor analysis, because the observed correlations are perfectly consistent with the simplest possible factor-analytic hypothesis--the hypothesis of a single common factor. Some other correlation matrix might not fit the hypothesis of a single common factor, but might fit the hypothesis of two or three or four common factors. The fewer factors the simpler the hypothesis. Since simple hypothesis generally have logical scientific priority over more complex hypotheses, hypotheses involving fewer factors are considered to be preferable to those involving more factors. That is, you accept at least tentatively the simplest hypothesis (i.e., involving the fewest factors) that is not clearly contradicted by the set of observed correlations. Like many writers, I'll let m denote the hypothesized number of common factors.

Without getting deeply into the mathematics, we can say that factor analysis attempts to express each variable as the sum of *common* and *unique* portions. The common portions of all the variables are by definition fully explained by the common factors, and the unique portions are ideally perfectly uncorrelated with each other. The degree to which a given data set fits this condition can be judged from an analysis of what is usually called the "residual correlation matrix".

The name of this matrix is somewhat misleading because the entries in the matrix are typically not correlations. If there is any doubt in your mind about some particular printout, look for the diagonal entries in the matrix, such as the "correlation" of the first variable with itself, the second with itself, etc. If these diagonal entries are not all exactly 1, then the matrix printed is not a correlation matrix. However, it can typically be transformed into a correlation matrix by dividing each off-diagonal entry by the square roots of the two corresponding diagonal entries. For instance, if the first two diagonal entries are .36 and .64, and the off-diagonal entry in position [1,2] is .3, then the residual correlation is $.3 / (.6 * .8) = 5/8 = .625$.

Correlations found in this way are the correlations that would have to be allowed among the "unique" portions of the variables in order to make the common portions of the variables fit the hypothesis of m common factors. If these calculated correlations are so high that they are inconsistent with the hypothesis that they are 0 in the population, then the hypothesis of m common factors is rejected. Increasing m always lowers these correlations, thus producing a hypothesis more consistent with the data.

We want to find the simplest hypothesis (that is, the lowest m) consistent with the data. In this respect, a factor analysis can be compared to episodes in scientific history that took decades or centuries to develop. Copernicus realized that the earth and other planets moved around the sun, but he first hypothesized that their orbits were circles. Kepler later realized that the orbits were better described as

ellipses. A circle is a simpler figure than an ellipse, so this episode of scientific history illustrates the general point that we start with a simple theory and gradually make it more complex to better fit the observed data.

The same principle can be observed in the history of experimental psychology. In the 1940s, experimental psychologists widely believed that all the basic principles of learning, that might even revolutionize educational practice, could be discovered by studying rats in mazes. Today that view is considered ridiculously oversimplified, but it does illustrate the general scientific point that it is reasonable to start with a simple theory and gradually move to more complex theories only when it becomes clear that the simple theory fails to fit the data.

This general scientific principle can be applied within a single factor analysis. Start with the simplest possible theory (usually $m = 1$), test the fit between that theory and the data, and then increase m as needed. Each increase in m produces a theory that is more complex but will fit the data better. Stop when you find a theory that fits the data adequately.

Each observed variable's *communality* is its estimated squared correlation with its own common portion--that is, the proportion of variance in that variable that is explained by the common factors. If you perform factor analyses with several different values of m , as suggested above, you will find that the communalities generally increase with m . But the communalities are not used to choose the final value of m . Low communalities are not interpreted as evidence that the data fail to fit the hypothesis, but merely as evidence that the variables analyzed have little in common with one another. Most factor analysis programs first estimate each variable's communality as the squared multiple correlation between that variable and the other variables in the analysis, then use an iterative procedure to gradually find a better estimate.

Factor analysis may use either correlations or *covariances*. The covariance cov_{jk} between two variables numbered j and k is their correlation times their two standard deviations: $\text{cov}_{jk} = r_{jk} s_j s_k$, where r_{jk} is their correlation and s_j and s_k are their standard deviations. A covariance has no very important substantive meaning, but it does have some very useful mathematical properties described in the next section. Since any variable correlates 1 with itself, any variable's covariance with itself is its variance--the square of its standard deviation. A correlation matrix can be thought of as a matrix of variances and covariances (more concisely, a covariance matrix) of a set of variables that have already been adjusted to standard deviations of 1. Therefore I shall often talk about a covariance matrix when we really mean either a correlation or covariance matrix. I will use R to denote either a correlation or covariance matrix of observed variables. This is admittedly awkward, but the matrix analyzed is nearly always a correlation matrix, and as explained later we need the letter C for the common-factor portion of R .

Matrix Decomposition and Rank

This *optional* section gives a little more detail on the mathematics of factor analysis. I assume you are familiar with the central theorem of analysis of variance: that the sum of squares of a dependent variable Y can be partitioned into components which sum to the total. In any analysis of variance the total sum of squares can be partitioned into model and residual components. In a two-way factorial analysis of variance with equal cell frequencies, the model sum of squares can be further partitioned into row, column, and interaction components.

The central theorem of factor analysis is that you can do something similar for an entire covariance matrix. A covariance matrix R can be partitioned into a common portion C which is explained by a set of factors, and a unique portion U unexplained by those factors. In matrix terminology, $R = C + U$, which means that each entry in matrix R is the sum of the corresponding entries in matrices C and U .

As in analysis of variance with equal cell frequencies, the explained component C can be broken down further. C can be decomposed into component matrices c_1, c_2, \dots , explained by individual factors.

Each of these one-factor components c_j equals the "outer product" of a column of "factor loadings".

The outer product of a column of numbers is the square matrix formed by letting entry jk in the matrix equal the product of entries j and k in the column. Thus if a column has entries .9, .8, .7, .6, .5, as in the earlier example, its outer product is

	.81	.72	.63	.54	.45
	.72	.64	.56	.48	.40
c_1	.63	.56	.49	.42	.35
	.54	.48	.42	.36	.30
	.45	.40	.35	.30	.25

Earlier I mentioned the off-diagonal entries in this matrix but not the diagonal entries. Each diagonal entry in a c_j matrix is actually the amount of variance in the corresponding variable explained by that factor. In our example, g correlates .9 with the first observed variable, so the amount of explained variance in that variable is $.9^2$ or .81, the first diagonal entry in this matrix.

In the example there is only one common factor, so matrix C for this example (denoted C_{55}) is $C_{55} = c_1$. Therefore the residual matrix U for this example (denoted U_{55}) is $U_{55} = R_{55} - c_1$. This gives the following matrix for U_{55} :

	.19	.00	.00	.00	.00
	.00	.36	.00	.00	.00
U_{55}	.00	.00	.51	.00	.00
	.00	.00	.00	.64	.00
	.00	.00	.00	.00	.75

This is the covariance matrix of the portions of the variables unexplained by the factor. As mentioned earlier, all off-diagonal entries in U_{55} are 0, and the diagonal entries are the amounts of unexplained or unique variance in each variable.

Often C is the sum of several matrices c_j , not just one as in this example. The number of c -matrices which sum to C is the *rank* of matrix C ; in this example the rank of C is 1. The rank of C is the number of common factors in that model. If you specify a certain number m of factors, a factor analysis program then derives two matrices C and U which sum to the original correlation or covariance matrix R , making the rank of C equal m . The larger you set m , the closer C will approximate R . If you set $m = p$, where p is the number of variables in the matrix, then every entry in C will exactly equal the corresponding entry in R , leaving U as a matrix of zeros. The idea is to see how low you can set m and still have C provide a reasonable approximation to R .

How Many Cases and Variables?

The clearer the true factor structure, the smaller the sample size needed to discover it. But it would be very difficult to discover even a very clear and simple factor structure with fewer than about 50 cases, and 100 or more cases would be much preferable for a less clear structure.

The rules about number of variables are very different for factor analysis than for regression. In factor analysis it is perfectly okay to have many more variables than cases. In fact, generally speaking the more variables the better, so long as the variables remain relevant to the underlying factors.

How Many Factors?

This section describes two rules for choosing the number of factors. Readers familiar with factor analysis will be surprised to find no mention of Kaiser's familiar eigenvalue rule or Cattell's scree test. Both rules are mentioned later, though as explained at that time I consider both rules obsolescent. Also both use eigenvalues, which I have not yet introduced.

Of the two rules that are discussed in this section, the first uses a formal significance test to identify the number of common factors. Let N denote the sample size, p the number of variables, and m the number of factors. Also R_U denotes the residual matrix U transformed into a correlation matrix, $|R_U|$ is its determinant, and $\ln(1/|R_U|)$ is the natural logarithm of the reciprocal of that determinant.

To apply this rule, first compute $G = N-1-(2p+5)/6-(2/3)m$. Then compute

$$\text{Chi-square} = G \ln(1/|R_U|)$$

with

$$df = .5[(p-m)^2 - p - m]$$

If it is difficult to compute $\ln(1/|R_U|)$, that expression is often well approximated by r_U^2 , where the summation denotes the sum of all squared correlations above the diagonal in matrix R_U .

To use this formula to choose the number of factors, start with $m = 1$ (or even with $m = 0$) and compute this test for successively increasing values of m , stopping when you find nonsignificance; that value of m is the smallest value of m that is not significantly contradicted by the data. The major difficulty with this rule is that in my experience, with moderately large samples it leads to more factors than can successfully be interpreted.

I recommend an alternative approach. This approach was once impractical, but today is well within reach. Perform factor analyses with various values of m , complete with rotation, and choose the one that gives the most appealing structure.

Rotation

In the opening example on curiosity, I mentioned individual factors that Rubenstein described: enjoyment of reading, interest in science, etc. Rotation is the step in factor analysis that allows you to identify meaningful factor names or descriptions like these.

Linear Functions of Predictors

To understand rotation, first consider a problem that doesn't involve factor analysis. Suppose you want to predict the grades of college students (all in the same college) in many different courses, from their scores on general "verbal" and "math" skill tests. To develop predictive formulas, you have a body of past data consisting of the grades of several hundred previous students in these courses, plus the scores of those students on the math and verbal tests. To predict grades for present and future students, you could use these data from past students to fit a series of two-variable multiple regressions, each regression predicting grade in one course from scores on the two skill tests.

Now suppose a co-worker suggests summing each student's verbal and math scores to obtain a composite "academic skill" score I'll call AS, and taking the difference between each student's verbal and math scores to obtain a second variable I'll call VMD (verbal-math difference). The co-worker

suggests running the same set of regressions to predict grades in individual courses, except using AS and VMD as predictors in each regression, instead of the original verbal and math scores. In this example, you would get exactly the same predictions of course grades from these two families of regressions: one predicting grades in individual courses from verbal and math scores, the other predicting the same grades from AS and VMD scores. In fact, you would get the same predictions if you formed composites of 3 math + 5 verbal and 5 verbal + 3 math, and ran a series of two-variable multiple regressions predicting grades from these two composites. These examples are all *linear functions* of the original verbal and math scores.

The central point is that if you have m predictor variables, and you replace the m original predictors by m linear functions of those predictors, you generally neither gain or lose any information--you could if you wish use the scores on the linear functions to reconstruct the scores on the original variables. But multiple regression uses whatever information you have in the optimum way (as measured by the sum of squared errors in the current sample) to predict a new variable (e.g. grades in a particular course). Since the linear functions contain the same information as the original variables, you get the same predictions as before.

Given that there are many ways to get exactly the same predictions, is there any advantage to using one set of linear functions rather than another? Yes there is; one set may be *simpler* than another. One particular pair of linear functions may enable many of the course grades to be predicted from just one variable (that is, one linear function) rather than from two. If we regard regressions with fewer predictor variables as simpler, then we can ask this question: Out of all the possible pairs of predictor variables that would give the same predictions, which is simplest to use, in the sense of minimizing the number of predictor variables needed in the typical regression? The pair of predictor variables maximizing some measure of simplicity could be said to have *simple structure*. In this example involving grades, you might be able to predict grades in some courses accurately from just a verbal test score, and predict grades in other courses accurately from just a math score. If so, then you would have achieved a "simpler structure" in your predictions than if you had used both tests for all predictions.

Simple Structure in Factor Analysis

The points of the previous section apply when the predictor variables are factors. Think of the m factors F as a set of independent or predictor variables, and think of the p observed variables X as a set of dependent or criterion variables. Consider a set of p multiple regressions, each predicting one of the variables from all m factors. The standardized coefficients in this set of regressions form a $p \times m$ matrix called the *factor loading matrix*. If we replaced the original factors by a set of linear functions of those factors, we would get exactly the same predictions as before, but the factor loading matrix would be different. Therefore we can ask which, of the many possible sets of linear functions we might use, produces the simplest factor loading matrix. Specifically we will define simplicity as the number of zeros or near-zero entries in the factor loading matrix--the more zeros, the simpler the structure. Rotation does not change matrix C or U at all, but does change the factor loading matrix.

In the extreme case of simple structure, each X -variable will have only one large entry, so that all the others can be ignored. But that would be a simpler structure than you would normally expect to achieve; after all, in the real world each variable isn't normally affected by only one other variable. You then name the factors subjectively, based on an inspection of their loadings.

In common factor analysis the process of rotation is actually somewhat more abstract than I have implied here, because you don't actually know the individual scores of cases on factors. However, the statistics for a multiple regression that are most relevant here--the multiple correlation and the standardized regression slopes--can all be calculated just from the correlations of the variables and

factors involved. Therefore we can base the calculations for rotation to simple structure on just those correlations, without using any individual scores.

A rotation which requires the factors to remain uncorrelated is an *orthogonal* rotation, while others are *oblique* rotations. Oblique rotations often achieve greater simple structure, though at the cost that you must also consider the matrix of factor intercorrelations when interpreting results. Manuals are generally clear which is which, but if there is ever any ambiguity, a simple rule is that if there is any ability to print out a matrix of factor correlations, then the rotation is oblique, since no such capacity is needed for orthogonal rotations.

An Example

Table 1 illustrates the outcome of rotation with a factor analysis of 24 measures of mental ability.

Table 1
Oblique Promax rotation of 4 factors of 24 mental ability variables
From Gorsuch (1983)

	Verbal	Numer- ical	Visual	Recog- nition
General information	.80	.10	-.01	-.06
Paragraph comprehension	.81	-.10	.02	.09
Sentence completion	.87	.04	.01	-.10
Word classification	.55	.12	.23	-.08
Word meaning	.87	-.11	-.01	.07
Add	.08	.86	-.30	.05
Code	.03	.52	-.09	.29
Counting groups of dots	-.16	.79	.14	-.09
Straight & curved capitals	-.01	.54	.41	-.16
Woody-McCall mixed	.24	.43	.00	.18
Visual perception	-.08	.03	.77	-.04
Cubes	-.07	-.02	.59	-.08
Paper form board	-.02	-.19	.68	-.02
Flags	.07	-.06	.66	-.12
Deduction	.25	-.11	.40	.20
Numerical puzzles	-.03	.35	.37	.06
Problem reasoning	.24	-.07	.36	.21
Series completion	.21	.05	.49	.06
Word recognition	.09	-.08	-.13	.66
Number recognition	-.04	-.09	-.02	.64
Figure recognition	-.16	-.13	.43	.47
Object-number	.00	.09	-.13	.69
Number-figure	-.22	.23	.25	.42
Figure-word	.00	.05	.15	.37

This table reveals quite a good simple structure. Within each of the four blocks of variables, the high values (above about .4 in absolute value) are generally all in a single column--a separate column for each of the four blocks. Further, the variables within each block all seem to measure the same general kind of mental ability. The major exception to both these generalizations comes in the third block. The variables in that block seem to include measures of both visual ability and reasoning, and the reasoning variables (the last four in the block) generally have loadings in column 3 not far above their loadings in

one or more other columns. This suggests that a 5-factor solution might be worth trying, in the hope that it might yield separate "visual" and "reasoning" factors. The factor names in Table 1 were given by Gorsuch, but inspection of the variables in the second block suggests that "simple repetitive tasks" might be a better name for factor 2 than "numerical".

I don't mean to imply that you should always try to make every variable load highly on only one factor. For instance, a test of ability to deal with arithmetic word problems might well load highly on both verbal and mathematical factors. This is actually one of the advantages of factor analysis over cluster analysis, since you cannot put the same variable in two different clusters.

Principal Component Analysis (PCA)

Basics

I have introduced principal component analysis (PCA) so late in this chapter primarily for pedagogical reasons. It solves a problem similar to the problem of common factor analysis, but different enough to lead to confusion. It is no accident that common factor analysis was invented by a scientist (differential psychologist Charles Spearman) while PCA was invented by a statistician. PCA states and then solves a well-defined statistical problem, and except for special cases always gives a unique solution with some very nice mathematical properties. One can even describe some very artificial practical problems for which PCA provides the exact solution. The difficulty comes in trying to relate PCA to real-life scientific problems; the match is simply not very good. Actually PCA often provides a good approximation to common factor analysis, but that feature is now unimportant since both methods are now easy enough.

The central concept in PCA is representation or summarization. Suppose we want to replace a large set of variables by a smaller set which best summarizes the larger set. For instance, suppose we have recorded the scores of hundreds of pupils on 30 mental tests, and we don't have the space to store all those scores. (This is a very artificial example in the computer age, but was more appealing before then, when PCA was invented.) For economy of storage we would like to reduce the set to 5 scores per pupil, from which we would like to be able to reconstruct the original 30 scores as accurately as possible.

Let p and m denote respectively the original and reduced number of variables--30 and 5 in the current example. The original variables are denoted X , the summarizing variables F for factor. In the simplest case our measure of accuracy of reconstruction is the sum of p squared multiple correlations between X -variables and the predictions of X made from the factors. In the more general case we can weight each squared multiple correlation by the variance of the corresponding X -variable. Since we can set those variances ourselves by multiplying scores on each variable by any constant we choose, this amounts to the ability to assign any weights we choose to the different variables.

We now have a problem which is well-defined in the mathematical sense: reduce p variables to a set of m linear functions of those variables which best summarize the original p in the sense just described. It turns out, however, that infinitely many linear functions provide equally good summaries. To narrow the problem to one unique solution, we introduce three conditions. First, the m derived linear functions must be mutually uncorrelated. Second, any set of m linear functions must include the functions for a smaller set. For instance, the best 4 linear functions must include the best 3, which include the best 2, which include the best one. Third, the squared weights defining each linear function must sum to 1. These three conditions provide, for most data sets, one unique solution. Typically there are p linear functions (called *principal components*) declining in importance; by using all p you get perfect

reconstruction of the original X-scores, and by using the first m (where m ranges from 1 to p) you get the best reconstruction possible for that value of m .

Define each component's *eigenvector* or *characteristic vector* or *latent vector* as the column of weights used to form it from the X-variables. If the original matrix R is a correlation matrix, define each component's *eigenvalue* or *characteristic value* or *latent value* as its sum of squared correlations with the X-variables. If R is a covariance matrix, define the eigenvalue as a weighted sum of squared correlations, with each correlation weighted by the variance of the corresponding X-variable. The sum of the eigenvalues always equals the sum of the diagonal entries in R.

Nonunique solutions arise only when two or more eigenvalues are exactly equal; it then turns out that the corresponding eigenvectors are not uniquely defined. This case rarely arises in practice, and I shall ignore it henceforth.

Each component's eigenvalue is called the "amount of variance" the component explains. The major reason for this is the eigenvalue's definition as a weighted sum of squared correlations. However, it also turns out that the actual variance of the component scores equals the eigenvalue. Thus in PCA the "factor variance" and "amount of variance the factor explains" are always equal. Therefore the two phrases are often used interchangeably, even though conceptually they stand for very different quantities.

The Number of Principal Components

It may happen that m principal components will explain all the variance in a set of X-variables--that is, allow perfect reconstruction of X--even though $m < p$. However, in the absence of this event, there is no significance test on the number of principal components. To see why, consider first a simpler problem: testing the null hypothesis that a correlation between two variables is 1.0. This hypothesis implies that all points in the population fall in a straight line. It then follows that all points in any sample from that population must also fall in a straight line. From that it follows that if the correlation is 1.0 in the population, it must also be 1.0 in every single sample from that population. Any deviation from 1.0, no matter how small, contradicts the null hypothesis. A similar argument applies to the hypothesis that a multiple correlation is 1.0. But the hypothesis that m components account for all the variance in p variables is essentially the hypothesis that when the variables are predicted from the components by multiple regression, the multiple correlations are all 1.0. Thus even the slightest failure to observe this in a sample contradicts the hypothesis concerning the population.

If the last paragraph's line of reasoning seems to contain a gap, it is in the failure to distinguish between sampling error and measurement error. Significance tests concern only sampling error, but it is reasonable to hypothesize that an observed correlation of, say, .8 differs from 1.0 only because of measurement error. However, the possibility of measurement error implies that you should be thinking in terms of a common factor model rather a component model, since measurement error implies that there is some variance in each X-variable not explained by the factors.

Eigenvalue-Based Rules for Selecting the Number of Factors

Henry Kaiser suggested a rule for selecting a number of factors m less than the number needed for perfect reconstruction: set m equal to the number of eigenvalues greater than 1. This rule is often used in common factor analysis as well as in PCA. Several lines of thought lead to Kaiser's rule, but the simplest is that since an eigenvalue is the amount of variance explained by one more factor, it doesn't make sense to add a factor that explains less variance than is contained in one variable. Since a component analysis is supposed to summarize a set of data, to use a component that explains less than a

variance of 1 is something like writing a summary of a book in which one section of the summary is longer than the book section it summarizes--which makes no sense. However, Kaiser's major justification for the rule was that it matched pretty well the ultimate rule of doing several factor analyses with different numbers of factors, and seeing which analysis made sense. That ultimate rule is much easier today than it was a generation ago, so Kaiser's rule seems obsolete.

An alternative method called the *scree test* was suggested by Raymond B. Cattell. In this method you plot the successive eigenvalues, and look for a spot in the plot where the plot abruptly levels out. Cattell named this test after the tapering "scree" or rockpile at the bottom of a landslide. One difficulty with the scree test is that it can lead to very different conclusions if you plot the square roots or the logarithms of the eigenvalues instead of the eigenvalues themselves, and it is not clear why the eigenvalues themselves are a better measure than these other values.

Another approach is very similar to the scree test, but relies more on calculation and less on graphs. For each eigenvalue L , define S as the sum of all later eigenvalues plus L itself. Then L/S is the proportion of previously-unexplained variance explained by L . For instance, suppose that in a problem with 7 variables the last 4 eigenvalues were .8, .2, .15, and .1. These sum to 1.25, so 1.25 is the amount of variance unexplained by a 3-factor model. But $.8/1.25 = .64$, so adding one more factor to the 3-factor model would explain 64% of previously-unexplained variance. A similar calculation for the fifth eigenvalue yields $.2/(.2+.15+.1) = .44$, so the fifth principal component explains only 44% of previously unexplained variance.

Some Relations Among Output Values

A number of relations exist among output values. Many people feel these relationships help them understand their output better. Others are just compulsive, and like to use these relations to confirm that gremlins have not attacked their computer program. The major relationships are the following:

1. Sum of eigenvalues = p
if the input matrix was a correlation matrix
Sum of eigenvalues = sum of input variances
if the input matrix was a covariance matrix
2. Proportion of variance explained = eigenvalue / sum of eigenvalues
3. Sum of squared factor loadings for j th principal component
= eigenvalue _{j}
4. Sum of squared factor loadings for variable i
= variance explained in variable i
= C_{ii} (diagonal entry i in matrix C)
= communality _{i} in common factor analysis
= variance of variable i if $m = p$
5. Sum of crossproducts between columns i and j of factor loading matrix
= C_{ij} (entry ij in matrix C)
6. The relations in #3, #4 and #5 are still true after rotation.
7. $R - C = U$. If necessary, rule 4 can be used to find the diagonal entries in C , then rule 7 can be used to find the diagonal entries in U .

Comparing Two Factor Analyses

Since factor loadings are among the most important pieces of output from a factor analysis, it seems natural to ask about the standard error of a factor loading, so that for instance we might test the significance of the difference between the factor loadings in two samples. Unfortunately, no very useful general formula for such a purpose can be derived, because of ambiguities in identifying the factors themselves. To see this, imagine that "math" and "verbal" factors explain roughly equal amounts of variance in a population. The math and verbal factors might emerge as factors 1 and 2 respectively in one sample, but in the opposite order in a second sample from the same population. Then if we mechanically compared, for instance, the two values of the loading of variable 5 on factor 1, we would actually be comparing variable 5's loading on the math factor to its loading on the verbal factor. More generally, it is never completely meaningful to say that one particular factor in one factor analysis "corresponds" to one factor in another factor analysis. Therefore we need a completely different approach to studying the similarities and differences between two factor analyses.

Actually, several different questions might be phrased as questions about the similarity of two factor analyses. First we must distinguish between two different data formats:

1. *Same variables, two groups.* The same set of measures might be taken on men and women, or on treatment and control groups. The question then arises whether the two factor structures are the same.
2. *One group, two conditions or two sets of variables.* Two test batteries might be given to a single group of subjects, and questions asked about how the two sets of scores differ. Or the same battery might be given under two different conditions.

The next two sections consider these questions separately.

Comparing Factor Analyses in Two Groups

In the case of two groups and one set of variables, a question about factor structure is obviously not asking whether the two groups differ in means; that would be a question for MANOVA (multivariate analysis of variance). Unless the two sets of means are equal or have somehow been made equal, the question is also not asking whether a correlation matrix can meaningfully be computed after pooling the two samples, since differences in means would destroy the meaning of such a matrix.

The question, "Do these two groups have the same factor structure?" is actually quite different from the question, "Do they have the same factors?" The latter question is closer to the question, "Do we need two different factor analyses for the two groups?" To see the point, imagine a problem with 5 "verbal" tests and 5 "math" tests. For simplicity imagine all correlations between the two sets of tests are exactly zero. Also for simplicity consider a component analysis, though the same point can be made concerning a common factor analysis. Now imagine that the correlations among the 5 verbal tests are all exactly .4 among women and .8 among men, while the correlations among the 5 math tests are all exactly .8 among women and .4 among men. Factor analyses in the two groups separately would yield different factor structures but identical factors; in each gender the analysis would identify a "verbal" factor which is an equally-weighted average of all verbal items with 0 weights for all math items, and a "math" factor with the opposite pattern. In this example nothing would be gained from using separate factor analyses for the two genders, even though the two factor structures are quite different.

Another important point about the two-group problem is that an analysis which derives 4 factors for group A and 4 for group B has as many factors total as an analysis which derives 8 in the combined group. Thus the practical question may be not whether analyses deriving m factors in each of two groups fit the data better than an analysis deriving m factors in the combined group. Rather the two

separate analyses should be compared to an analysis deriving $2m$ factors in the combined group. To make this comparison for component analysis, sum the first m eigenvalues in each separate group, and compare the mean of those two sums to the sum of the first $2m$ eigenvalues in the combined group. It would be very rare that this analysis suggests that it would be better to do separate factor analyses for the two groups. This same analysis should give at least an approximate answer to the question for common factor analysis as well.

Suppose the question really is whether the two factor structures are identical. This question is very similar to the question as to whether the two correlation or covariance matrices are identical--a question which is precisely defined with no reference to factor analysis at all. Tests of these hypotheses are beyond the scope of this work, but a test on the equality of two covariance matrices appears in Morrison (1990) and other works on multivariate analysis.

Comparing Factor Analyses of Two Sets of Variables in a Single Group

One question people often ask is whether they should analyze variable sets A and B together or separately. My answer is usually "together", unless there is obviously no overlap between the two domains studied. After all, if the two sets of variables really are unrelated then the factor analysis will tell you so, deriving one set of factors for set A and another for set B. Thus to analyze the two sets separately is to prejudge part of the very question the factor analysis is supposed to answer for you.

As in the case of two separate samples of cases, there is a question which often gets phrased in terms of factors but which is better phrased as a question about the equality of two correlation or covariance matrices--a question which can be answered with no reference to factor analysis. In the present instance we have two parallel sets of variables; that is, each variable in set A parallels one in set B. In fact, sets A and B may be the very same measures administered under two different conditions. The question then is whether the two correlation matrices or covariance matrices are identical. This question has nothing to do with factor analysis, but it also has little to do with the question of whether the AB correlations are high. The two correlation or covariance matrices within sets A and B might be equal regardless of whether the AB correlations are high or low.

Darlington, Weinberg, and Walberg (1973) described a test of the null hypothesis that the covariance matrices for variable sets A and B are equal when sets A and B are measured in the same sample of cases. It requires the assumption that the AB covariance matrix is symmetric. Thus for instance if sets A and B are the same set of tests administered in years 1 and 2, the assumption requires that the covariance between test X in year 1 and test Y in year 2 equal the covariance between test X in year 2 and test Y in year 1. Given this assumption, You can simply form two sets of scores I'll call A+B and A-B, consisting of the sums and differences of parallel variables in the two sets. It then turns out that the original null hypothesis is equivalent to the hypothesis that all the variables in set A+B are uncorrelated with all variables in set A-B. This hypothesis can be tested with MANOVA.

REFERENCES

- Darlington, Richard B., Sharon Weinberg, and Herbert Walberg (1973). Canonical variate analysis and related techniques. *Review of Educational Research*, 453-454.
- Gorsuch, Richard L. (1983) *Factor Analysis*. Hillsdale, NJ: Erlbaum
- Morrison, Donald F. (1990) *Multivariate Statistical Methods*. New York: McGraw-Hill.
- Rubenstein, Amy S. (1986). An item-level analysis of questionnaire-type measures of intellectual curiosity. Cornell University Ph. D. thesis.