5 The Problem of Factor Scores

5.1. FACTOR ANALYSIS AS A TEST CONSTRUCTION DEVICE

It might seem that a factor-analytic study would be painfully incomplete if we did not obtain some assessment of the common factors themselves (i.e., of the common factor scores, or factor values, of the individual subjects in the study) as well as estimates of factor loadings. One reason why a factor-analytic study can give the appearance of leading nowhere while it actually leads somewhere is that it may serve as a guide to the construction of homogeneous tests of the traits or generic properties identified in the analysis. Because there may be a time gap between the publication of the factor analysis and the appearance of the consequent invented test, we may fail to notice the follow-up work that adds a practical justification to the earlier study.

For example, Thurstone's work yielding a set of correlated "primary mental abilities" led, through a series of refinements of sets of items, to a battery of standardized and easily usable tests of those abilities. The simple sum scores (total number of right answers—the "number right" score) on these developed cognitive tests can reasonably be thought of as the ultimate result of the factor-analytic work and as good "practical" measures of the factors identified. In the case of Thurstone's primary mental abilities, it is an open question whether the outcome of the whole enterprise was worthwhile. McNemar[1] (as mentioned previously) has surveyed evidence in support of a conclusion that the measures of general intelligence favored in Britain that stem naturally from the earliest factor-

[1]See McNemar (1964).

analytic tradition prove more effective in further research than the primary abilities that stem from Thurstone's oblique simple structure notions. If this issue remains controversial, it still serves to show that the effectiveness of research based on mental tests can rest on the effectiveness of the prior analysis on which the test construction was based.

The phrase "construction of homogeneous tests" was used as though its meaning is already quite obvious. Intuitively, we would expect that a set of measures that were indicators of just one generic property that they shared in common would be scored collectively to give a single measure of that property, and we would regard them as a *homogeneous* set of indicators (homogeneous = of the same kind). In the context of psychological or educational testing, such a combination of measures is said to yield a homogeneous test. The theory of tests and test construction tends to be dealt with rather separately from the theory of factor analysis. One reason for this lies in doubts as to whether the common factor model can be applied to dichotomous items. Crude practical devices have been developed for selecting and combining items intended to form homogeneous tests, and there have been some conceptual confusions about the meaning and the assessment of test homogeneity. But it seems reasonable to recognize that the primary theoretical conception of a homogeneous test (or set of measures) is one whose items measure only one (generic) property in common. When that property is held constant, the measures are statistically independent and certainly uncorrelated. In the context of this kind of thinking, we call the generic property a *latent trait*. The view taken here is that common factors and latent traits are essentially the same quantities—they are quantities that explain the relations among our measures. However, dichotomous items do not meet the usual assumptions of common factor analysis (e.g., we cannot assume their regressions on the factors are linear). Hence theory for items has, regrettably, been developed separately from factor theory, and their essential unity is not always recognized. More will be said about this in Chapter 7. The point here is that factor analysis, or closely related techniques, does provide a very good way to construct homogeneous tests of interesting generic properties of the subjects we study, though its role as a test construction device is not always recognized.

To repeat the central point of this section, although the connections are not always obvious from the literature, a usual sequel to factor analysis is the construction of a test intended to measure a common factor that has been "identified," "discovered," or "invented." Where such tests are produced, we do not need factor scoring procedures as such.

5.2. THE ESTIMATION OF COMMON FACTORS

Suppose that we know the population parameters of the common factor model in some application of it. That is, we know the nm factor loadings in the factor

pattern and the n residual variances, and, if the factors are correlated, we know their correlations. We can draw a subject from the population and measure his or her values of the n variables. The question is, what can we say about the m factor "scores" of the individual, corresponding to the n observed scores? It should be immediately clear that the factor scores are not uniquely determined by the observed scores. This fact has worried some workers in the field and is sometimes perceived as a basic fault of the model, justifying its abandonment in favor of, for example, principal components or images (Section 2.3). The position taken here is that we should not expect factor scores, measures of generic properties, to be exactly determined by means of a small number of empirical measures, each with its specific property and error of measurement, any more than we expect a population parameter to be precisely determined by statistics from a small sample of subjects. If the measures are drawn from a well-defined, appropriate behavior domain, they tend to yield estimates of factors that become increasingly precise as the number of measures is increased, just as the subjects drawn from a well-defined population tend to yield estimates of the model parameters that become increasingly precise as the number of subjects is increased. If we adopt a conception of our field of inquiry as a behavior domain, we may think of the factor score as a limit of its estimate as we draw all possible variables from the behavior domain. (See Section 5.2.)

To return to our question: We know the regression weights of the n observed variables on the m common factors, and we assumed at the outset that the regressions are linear. It is natural to ask if we can now reverse the roles of the observed variables and the factors. That is, we wish to estimate the factors by determining their regressions on the observed variables. We want m linear combinations $\hat{x}_p$ of the n observed variables that will have maximum correlations with the unknown factors in the population, or, with the right scaling of variables and factors, we want m linear combinations $\hat{x}_p$ whose discrepancies $(x_p - \hat{x}_p)$ from the unknown factors x_p in the population are as small as possible (i.e., have minimum variance). Let us write

$$\hat{x}_p = \sum_{j=1}^{n} b_{pj} y_j \qquad p = 1, \ldots, m \tag{5.2.1}$$

to represent a general expression for the *regression estimates* $\hat{x}_p$ of the factor scores x_p. We need a procedure for computing the factor-variable (f–v) regression weights b_{pj} (i.e., the regression weights of the factors on the tests). The f–v regression weights, b_{pj}, can be computed by the standard formulas of regression theory from knowledge of the correlations between the factors and the variables and the correlations between the variables. As usual we assume that a computer program is looking after this problem for us. The f–v regression weights are always necessarily different from the v–f regression weights (i.e., the factor

loadings). (They would be the same theoretically if the observed variables were uncorrelated, but then all the regression weights would be zero; there would be no common factors, and the whole case would have degenerated into triviality.) We can arrange the f–v regression weights b_{pj} in either an $(m \times n)$ matrix as the order of its coefficients suggests or transpose it as is usually convenient and have the computer print it out as an $(n \times m)$ matrix, of which the pth column contains the n regression weights of the pth factor on the n variables. This matrix would be unambiguously recognizable in computer printout or in a program write up, if it is described as "regression weights for estimating factors," "weights for least-squares estimation of factor scores," or a clearly equivalent phrase. It is desirable to have in the printout also the multiple correlation between each factor and the n variables. This is a fundamental quantity, as we can think of it as an index of the precision of the estimates. We would not feel that we had used enough measures to estimate their generic property if this coefficient were low. (See following for typical coefficients). Table 5.2.1 gives the f–v regression weights calculated for the Thurstone case, from (1) the orthogonal factor pattern in Table 2.2.6; and (2) the oblique factor pattern in Table 3.2.2 and also gives the multiple correlation of each factor with the variables. The estimates $\hat{x}_p$ are

TABLE 5.2.1
Regression Weights for Thurstone Case

(a) From Orthogonal Pattern

	I	II	III
1	.451	-.135	-.028
2	.485	-.013	-.133
3	.263	-.025	-.054
4	-.152	.572	-.072
5	-.110	.347	.013
6	-.019	.224	-.072
7	-.126	-.139	.650
8	.027	-.072	.197
9	-.095	.023	.286

(b) From Oblique Pattern

	I	II	III
1	.353	.044	.064
2	.391	.049	.071
3	.226	.028	.041
4	.025	.406	.056
5	.020	.319	.044
6	.012	.203	.028
7	.025	.041	.359
8	.019	.031	.269
9	.017	.029	.250

(c) Correlation Matrix of Estimates From Orthogonal Pattern

	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$
$\hat{x}_1$	1.000	-.072	-.069
$\hat{x}_2$	-.072	1.000	-.098
$\hat{x}_3$	-.069	-.098	1.000

(d) Cross-Correlation Matrix of Estimates and Factors From Orthogonal Pattern

	x_1	x_2	x_3
$\hat{x}_1$	.927	-.062	-.059
$\hat{x}_2$	.066	.872	-.084
$\hat{x}_3$	-.064	-.086	.855

computed by combining the standard scores of each subject, using (5.2.1), with these regression coefficients.

The mathematical theory of the problem reveals the following further properties of regression estimators:

1. By definition, each estimator $\hat{x}_p$ is uncorrelated with its own residual $x_p - \hat{x}_p$, but also each estimator $\hat{x}_p$ is uncorrelated with the residual $x_q - \hat{x}_q$ of every other estimate.

2. In general, when the common factors x_p in the model are uncorrelated (i.e., when we have used the orthogonal model) the estimates $\hat{x}_p$ are mutually correlated. The correlation matrix of the estimates for the Thurstone matrix is given in Table 5.2.1.

3. In general, when the common factors x_p are uncorrelated, the estimator $\hat{x}_p$ of one factor can be correlated with the other $m-1$ factors x_q $(q \neq p)$. The cross-correlation matrix of estimators and factors for the Thurstone case is shown in Table 5.2.1.

4. The regression estimators are biased, in the sense that *if* we could select a subpopulation of subjects all having the same factor score x_p, the mean of the estimator for the selected subpopulation would not be the same as the factor score on the basis of which they were selected. This causes a worry that if we had groups of subjects selected on the basis of experimental treatments and we wished to compare their mean factor scores, we *might* be misled by a comparison of means of regression estimators.

The properties 2 and 3 are rightly thought of as unfortunate defects of the regression estimators. In practice, these correlations tend to be low, so perhaps only theoretical purists should worry about them.

In practice and as already illustrated, we do not know the parameters of the model but have to estimate them. In practice, therefore, we use estimated correlations in the formulas for regression weights in place of the population values.

By definition, the regression estimators yield best estimators in the sense of least squares and have maximum correlations with the factors in the given population. They are based on one clear notion of a "best" choice for the population as a whole.

An alternative to regression estimation arises when we consider just one subject drawn from the population, not necessarily at random, and we want "best" estimates of just that subject's factor scores. We have no concern now with any other individual and certainly not with the population as a whole. There are two recognized ways to get best estimates of just one individual's factor scores, which yield the same answer so it can be encountered under two distinct headings. The first way is to get maximum likelihood (ML) estimates of the factor scores. That is, given the model itself, which states that, for this and other individuals,

$$y_j = \sum_{p=1}^{m} f_{jp}x_p + e_j \qquad j = 1, \ldots, n$$

we will assume that the subject's n unique factors $e_1, \ldots, e_n$ are normally distributed variables and choose values of $x_1, \ldots, x_m$ that make the obtained values of $y_1, \ldots, y_n$ maximally probable. The second way is known in the literature as *weighted least squares*, but to avoid ambiguity (as this sounds like a variation of the regression estimates) and for another reason it would be better to call it specifically *weighted least-squared residuals* (WLSR). We choose values of $x_1, \ldots, x_m$ for our individual to minimize the quantity

$$\phi = \sum_{j=1}^{n} \frac{e_j^2}{u_j^2} \tag{5.2.2}$$

where

$$e_j = y_j - \sum_{j=1}^{m} f_{jp}x_p \tag{5.2.3}$$

That is, we choose numbers $x_1, \ldots, x_m$ that minimize the sum of the ratios of the squares of the n unique scores to the variances of those unique scores. (This is what is meant by *weighted least-squared residuals*.)

The mathematician, on being presented with the task of finding an expression for the ML estimator of our individual's factor scores and with the task of finding an expression for the WLSR estimator, announces that the same solution applies to both problems. Again the problem yields an expression for a set of coefficients, let us say t_{pj}, calculated from knowledge of the factor loadings and residual variances from which we shall compute the ML/WLSR estimators, denoted by $\tilde{x}_1, \ldots, \tilde{x}_m$, as

$$\tilde{x}_p = \sum_{j=1}^{n} t_{pj}y_j. \tag{5.2.4}$$

[Note that we use a caret (ˆ) for the regression estimators and a tilde (˜) for the ML/WLSR estimators.]

The expression for ML/WLSR estimates was first given by Bartlett, and these estimates are often referred to in the literature as the *Bartlett estimates* of factor scores. We shall call them ML/WLSR estimates. As in the case of the regression estimates, we can expect to find the weights t_{pj} printed out by a computer program either as an $(m \times n)$ matrix or transposed into an $(n \times m)$ matrix. Unambiguous titles or descriptions would include "weights for maximum like-

lihood/weighted least-squares estimators of factor scores'' and ''Bartlett scoring weights.'' Table 5.2.2 gives the ML weights for the Thurstone case (from Tables 2.2.2 and 3.2.2).

Although we have discussed this estimation method as though it is for just one individual, obviously we can apply it to each of many individuals drawn from the population, and we can in fact ask what properties the ML estimates have in the population. We find the following:

1. Necessarily, they have lower correlations with the common factors than do the regression estimates. In practice, the difference would usually be small.
2. Like the regression estimates, the ML estimates are correlated with each other even when the factors themselves are uncorrelated.
3. Unlike the regression estimates, the ML estimates have zero correlations with noncorresponding factors. Thus they are unambiguously related to ''the right'' factors and unrelated to ''the wrong'' factors.
4. Unlike the regression estimates, the ML estimates are unbiased. That is, if

TABLE 5.2.2
ML/WLSR Weights for Thurstone Case

	(a) From Orthogonal Pattern			(b) From Oblique Pattern		
	I	II	III	I	II	III
1	.546	-.222	-.064	.405	.0	.0
2	.553	-.044	-.228	.449	.0	.0
3	.317	-.052	-.097	.260	.0	.0
4	-.229	.790	-.167	.0	.551	.0
5	-.165	.473	-.020	.0	.434	.0
6	-.038	.312	-.130	.0	.275	.0
7	-.198	-.266	.937	.0	.0	.553
8	.019	-.127	.282	.0	.0	.414
9	.142	-.001	.404	.0	.0	.385

(c) Correlation Matrix of Estimates from Orthogonal Pattern

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$
$\tilde{x}_1$	1.000	.079	.007
$\tilde{x}_2$	.079	1.000	.104
x_3	.077	.104	1.000

(d) Cross-Correlation Matrix of Estimates and Factors from Orthogonal Pattern

	x_1	x_2	x_3
$\tilde{x}_1$	.992		
$\tilde{x}_2$		.865	
x_3			.848

we could select a subpopulation of subjects all having the same factor score x_p, the mean of the estimator over the subpopulation selected would be x_p. It rather seems that properties 3 and 4 of ML estimators, compared with the regression estimates, give the advantage to the ML estimates for both individual purposes and research on groups. Given an individual drawn from the population, certainly we would use the ML estimator, and we can use the normal distribution to put confidence bounds on the true values in the usual way. For group comparisons, unambiguous and unbiased estimates would generally seem desirable.

A small technical point should be noted in passing. By the usual principles of regression theory, the regression estimation procedure divides the factor score x_p into two uncorrelated parts, the regression part $\hat{x}_p$ and the residual $x_p - \hat{x}_p$. That is, we have

$$x_p = \hat{x}_p + d_p \tag{5.2.5}$$

say, where

$$d_p = x_p - \hat{x}_p \tag{5.2.6}$$

and $\hat{x}_p$ and d_p are uncorrelated. It turns out that the ML estimator $\tilde{x}_p$ itself can be written as the sum of the factor score x_p and a discrepancy that are uncorrelated. That is, we have

$$\tilde{x}_p = x_p + \delta_p \tag{5.2.7}$$

say, where

$$\delta_p = \tilde{x}_p - x_p \tag{5.2.8}$$

and x_p and δ_p are uncorrelated. The result is that the variance of $\tilde{x}_p$ is the sum of the (unit) variance of x_p and the variance of δ_p, the error about the true value. If the computer program prints out the variances of the ML estimators, these must be greater than one, and we can compute the standard deviation of the error term and hence get confidence bounds. If the computer program prints out the variances of the regression estimators, these must be less than one, and of course we cannot obtain confidence bounds, which are meaningless for biased estimators. This paragraph is a technical aside that can be ignored, except that the reader should note the implied device for deciding whether given factor score estimates, inadequately labeled, are regression estimates or ML estimates. If their variances are less than one, they are regression scores. If their variances are greater than one, they are ML scores. To apply this test, it is necessary to know that they are one of these two, however.

Other estimators have been described in the literature but do not seem to have anything to recommend them. Package programs often contain estimates that cannot be recommended without enough information for one to be able to tell

what device is being employed. In particular, the factor pattern itself, the v–f regression weight matrix, is sometimes used as though it were the f–v regression weight matrix. There is absolutely no foundation for this procedure in theory. It does not produce nonsensical results in general, however. In multivariate statistical methods, crude weighting methods have a way of giving results that are not at all horrible in comparison with optimal methods.

From one point of view, we could describe the process discussed in Section 5.1 of developing a test out of factor-analytic work as the process of assigning weights to a set of items equal to $+1$, -1, or 0, according to whether their factor loadings are "high positive," "high negative," or "low." Such crude but convenient scoring systems tend to yield sums of variables that are very highly correlated with combinations of them that employ "best" weights, in some precise sense of the word *best*, so it often may not seem worthwhile to work with optimal weights.

We can gain a sense of the typical numerical properties of factor score estimates by considering the special case of just one factor with equal factor loadings yielding equally correlated variables. In this case, the square of the correlation between the factor and its estimate from the n tests, whether regression or ML, is given by

$$\rho^2(x, \hat{x}) = \frac{nr}{1 + (n - 1)r} \tag{5.2.9}$$

where r is the correlation between any two variables y_j, y_k. (This expression is the same as the Spearman–Brown formula for the effect of test length on reliability. Here it is just a special case of factor score estimation theory. Whether or not true scores are factor scores depends on nonmathematical considerations.) By prevailing standards, in social science research, we might feel content with a correlation of about .85 or more between our estimate and that which we are estimating. By such a standard, it seems that 5 variables whose average correlation is above .4 or 2 with a correlation above .6 or 10 with an average above .2 will serve to determine a factor adequately.

The results on regression estimates and ML/WLSR estimates of common factor scores generalize, though not without complications, to models for linear structural relations as discussed in Chapter 4. In such models, if we wish to estimate the latent variables from the corresponding observed variables, we can do so with expressions that are of the same form as the ones used in the common factor model. However, in order to apply them it is necessary to do some extra manipulations of the parameter values obtained by fitting the model in order to compute the residual covariances of the observed variables about their regressions on the unobserved variables, to get ML/WLSR estimates, and to compute the covariances of the unobserved variables in order to get the regression estimates. A description of these manipulations without the language of matrix algebra would be quite uninformative.[2]

5.3. THE INDETERMINACY OF COMMON FACTORS

A number of investigators working on the mathematical theory of factor analysis have become convinced that common factor scores are seriously indeterminate quantities. As shown in equation (5.2.5), the unknown factor score x_p is the sum of its computable estimate $\hat{x}_p$ and the unknown residual d_p. It turns out that we can always invent arbitrary numbers d_p to add to $\hat{x}_p$ that yield arbitrary numbers x_p that have all the required properties of factor scores. There has been disagreement about the interpretation of the arbitrariness of the numbers x_p. We can work out the correlation between two sets of possible factor scores (estimates plus or minus invented numbers) that are chosen to be as dissimilar as possible. (This is a purely mathematical exercise, given the factor loadings.) The minimum possible correlation between alternative factor scores is given by $2\rho^2(x, \hat{x}) - 1$, where, as before, $\rho^2(x, \hat{x})$ is the square of the multiple correlation between the factor x and the n tests. If $\rho(x, \hat{x})$ is $1/\sqrt{2}$ (approximately .707), which does not seem a very low correlation, then $\rho^2(x, \hat{x})$ is $\frac{1}{2}$, and the correlation between the most dissimilar alternative factor scores that we can arbitrarily construct is zero.

The implications of this mathematical result are perhaps not yet fully understood and are still subject to disagreement. Some investigators have taken it to mean that the common factor model is subject to such a serious indeterminacy in its fundamental measurements that it should not be used, even if we have no interest in the factor scores themselves. The notion seems to be that if the factor scores of a set of examinees are not well determined by their scores on a set of tests, then the abstractive attributes that these scores serve to measure are not well defined by the characteristics of the tests in the set. That is, if the scores on a small number of items measuring a common property do not yield a unique score for the property, then correlatively the common features of the items do not provide a unique interpretation of the common property itself. Indeed, there is a sense in which this can be true.

As implied earlier, if we can imagine that the tests in a factor analysis are drawn from an infinite set of tests comprising a behavior domain, in which every

[2]See McDonald and Burr (1967) for a review of these and further results on factor score estimation. The importance of these results is probably diminishing as new, more general, models for linear structural analysis are developed. In research work on groups of examinees, we would now be likely to incorporate hypotheses about mean factor scores in models for simultaneous analysis in several populations (Section 6.2) or for repeated-measures, multimode data (Section 6.3), with no need to estimate individual factor scores and compare the means of the estimates across groups of examinees or across conditions.

common factor has infinitely many tests with nonzero loadings on it, then in the domain the common factor scores are correlated unity with their estimates from the infinity of tests. Suppose two investigators independently draw nonoverlapping subsets of tests from what is understood to be this behavior domain. Then the correlation between their factor score estimates will be equal to the product of the correlations of their estimates with the factor scores, as defined by the entire behavior domain. This correlation must be positive. As each augments the given set of tests to improve their estimates of the factor scores, the correlation between their factor score estimates must increase until in the limit the estimates coincide with each other and with the factor scores uniquely defined by the behavior domain they are both, by agreement, drawing from.

On the other hand, suppose that two investigators were to begin with the same set of tests, already factor analyzed, but with no idea of a defined behavior domain to draw from. Each then independently chooses further tests to add to the initial set to improve the estimation of the factors but with no concept of a defined behavior domain to draw them from. The augmented sets of tests are subject only to the requirement that the new tests have nonzero loadings on the same factors as the initial set. In such a case, there is no mathematical or logical reason why the two investigators should improve the agreement in their factor score estimates as they increase the number of their tests. Indeed, as this number becomes very large, the correlation between their estimates can be anything between unity and the quantity $2\rho^2(x, \hat{x}) - 1$, the minimum correlation between arbitrary mathematical constructions of factor scores, where $\rho^2(x, \hat{x})$ is calculated on the basis of the original, perhaps quite small, set of tests that both investigators started from. If the squared multiple correlation between the factor and the original tests is less than a half, then the correlation between their estimates can become and remain negative.

The mathematical theory just summarized in English is based on an extreme idealization of the process of inventing usable tests. Such idealizations are common in fields like classical physics, where the behavior of infinite homogeneous entities is commonly worked out as a theoretical approximation to the behavior of finite inhomogeneous entities. For the relation in our case between theory and practice, we must first make a further examination of the concept of a behavior domain.

We can imagine an exploratory factor analysis of a given, extremely large and thus virtually infinite collection of tests, whose extension (the range of its members) is defined by simple enumeration of its contents. In it, each factor score is determined almost certainly by scores on extremely large subsets of the tests. Given the tests and the factor analysis, the factor scores are measures of factor attributes that are definable *post facto* by abstraction of the common properties of the tests in those subsets. However, the collection of tests has not been supposed to have a clear denotation, a set of defining characteristics that distinguish tests that belong (and should belong) to the collection from those that do not (and

should not). This means that two investigators cannot be conceived of as independently drawing tests from this one collection, except in the literal sense that they share a list of all the tests and agree to choose tests out of that list. This will not happen in practice, because there is just no reason why they should wish to do such a thing. And if there is no agreed list, there is no reason, as we have seen, why the investigators should approach agreement with each other.

On the other hand, if a behavior domain is a set of tests with a stated denotation of their attributes, enabling us to distinguish tests that have these attributes from tests that do not, then two investigators can be conceived of as drawing tests from this one behavior domain whenever they invent a set of tests that possess the required denotation. We can then reasonably hope that as they augment their sets of tests to improve the measures of their defined common attributes, measures that in this case follow naturally from a confirmatory rather than an exploratory factor analysis, they will approach agreement in their measurements of these. The question of disagreement about the "interpretation" of the factors cannot arise as such in this case.

The conclusion we draw is that common factor scores appear to be centrally and essentially defined on the basis of the generalizability of the tests we use to tests we have not used that are in a clear sense of the same kind, in the sense that factor score estimates from the tests we have used are estimates of the scores defined by all the tests of that kind. As a special case of this, in classical test theory, the score on a test of finite length estimates the "true score" that would be obtained by augmenting the test to make one of infinite length. But unless the items in the imagined test of infinite length have a clear denotation in terms of their content, we can in theory find more than one test of infinite length that contains a given finite test and more than one "true score" that it is estimating.

In practice, the idealized theory described here can fail to approximate reality well for a number of reasons. There can be and no doubt will be hidden ambiguities in the denotation of the behavior domain, leading to distinct realizations of it in the constructed tests (e.g., hidden ambiguities in the concept of *extraversion* or of *clinical anxiety*). Also, it can be difficult or impossible to find many exemplars of a concept that do not form groups on the basis of other characteristics, which cause the number of common factors to multiply rapidly.

Whatever the difficulties of implementing behavior domain concepts in practice, we can use these concepts as a framework for examining a common alternative view of the problem of factor score indeterminacy. Some of those who regard this problem as indicating a serious flaw in the common factor model have suggested that we abandon the model in favor of other methods of analysis that yield very similar results, yet with all their quantities ("loadings" and "scores") uniquely determined by the test scores even from a quite small number of tests. For example, as shown in Chapter 2, principal component theory and image analysis give close approximations to common factor loadings. They also give close approximations to common factor score estimates. It has

therefore been argued that, for example, principal component scores are preferable to common factor scores because they are determinate and known and also that the methods are preferable to common factor analysis because they contain no indeterminate, unknown quantities. This argument, however, does not take into account the complementary facts that (1) the common factor score estimates are also determinate, so there is no reason to substitute principal component scores, for example, for estimates of common factor scores; (2) the principal component scores cannot have higher correlations with the corresponding principal component scores in a defined behavior domain from which the tests are drawn than do common factor score estimates with their corresponding factor scores in the domain. That is, principal components, images, and the like suffer a greater problem of indeterminacy than do common factors, in the sense that they have lower correlations with their counterparts in a defined behavior domain.

Some writers assert, then, that the common factor model has a serious indeterminacy problem in respect to its factor scores. Some further suggest that we should therefore use component theory, image theory, or even ad hoc adaptations of multidimensional scaling to the analysis of correlation coefficients as substitutes for the common factor model to achieve its intended purpose. In the present state of knowledge, the reader need not feel coerced by these assertions. Such alternative devices may be useful for certain purposes, but they have not been shown to be improvements on common factor analysis, preferably in its confirmatory form, for the purpose of investigating the generic properties of tests.[3]

5.4. MATHEMATICAL NOTES ON CHAPTER 5

By the theory of regression already given in Section 1.6, given the $(n \times m)$ matrix of regression weights $\mathbf{F}$ of the observed variables $\mathbf{y}$ on the factors $\mathbf{x}$, the $(m \times m)$ correlation matrix of the factors, $\mathbf{P}$, and the $(n \times n)$ correlation matrix of the observed variables, $\mathbf{R}$, we know that the correlation matrix $\mathbf{S}$, of order $(n \times m)$, of $\mathbf{y}$ and $\mathbf{x}$ is given by

$$\mathbf{S} = \mathbf{FP}.$$

Then by (1.6.16), applied to the present problem, the vector of regression estimates $\hat{\mathbf{x}}$ of $\mathbf{x}$, required in (5.2.1), is given by

$$\hat{\mathbf{x}} = \mathbf{B}'\mathbf{y} \tag{5.4.1}$$

[3] In a penetrating article, Guttman (1955) extended earlier results of Kestelman and used them to raise the problem discussed in this section. Mulaik and McDonald (1978) and McDonald (1977) give technical discussions of the issue, and McDonald and Mulaik (1979) give a nontechnical review of the question.

where

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{S} \tag{5.4.2}$$

or, alternatively,

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{FP}. \tag{5.4.3}$$

In the special case of uncorrelated factors, (5.4.3) reduces to

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{F}. \tag{5.4.4}$$

To obtain the ML/WLSR estimates $\hat{\mathbf{x}}$, we minimize the quantity

$$\phi = (\mathbf{y} - \mathbf{Fx})'\mathbf{U}^{-2}(\mathbf{y} - \mathbf{Fx}) \tag{5.4.5}$$

where, as usual, $\mathbf{U}^2$ is the diagonal matrix of uniquenesses. The idea is to minimize the sum of squares of an individual's residuals but weighted proportionally to the variance of each residual in the population. The alternative approach via maximum likelihood leads us to maximize

$$\phi^* = \frac{1}{(2\pi|\mathbf{U}^2|)^{1/2}} \exp -\frac{1}{2}(\mathbf{y} - \mathbf{Fx})'\mathbf{U}^{-2}(\mathbf{y} - \mathbf{Fx}) \tag{5.4.6}$$

and, by inspection, the two problems must have the same solution. The ML/WLSR estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$ turns out to be

$$\hat{\mathbf{x}} = (\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1}\mathbf{F}'\mathbf{U}^{-2}\mathbf{y} \tag{5.4.7}$$

a result obtained by methods outside the scope of the algebra introduced in Appendix A1.

In a computer program designed to produce factor scores after completing estimation of the factor pattern and so on, we would expect to find that the observed scores, formed into an $(n \times N)$ matrix $\mathbf{Y}$, are put into standard measure in the sample and stored on scratch tape while the main computations are going on; then, after $\mathbf{F}$ and $\mathbf{U}$ have been estimated, the estimates are employed to obtain the matrices required by (5.4.3) or (5.4.7). Finally, the scores on scratch tape would be called in, one subject at a time, and the estimates computed.

For the rest of these remarks we assume the model with uncorrelated factors. From (5.4.1) we find that

$$\begin{aligned} \mathrm{E}\{\hat{\mathbf{x}}\hat{\mathbf{x}}'\} &= \mathbf{B}'\mathrm{E}\{\mathbf{yy}'\}\mathbf{B} \\ &= \mathbf{B}'\mathbf{RB} \end{aligned} \tag{5.4.8}$$

hence, with (5.4.4)

$$\mathrm{E}\{\hat{\mathbf{x}}\hat{\mathbf{x}}'\} = \mathbf{F}'\mathbf{R}^{-1}\mathbf{F} \tag{5.4.9}$$

which in general is not a diagonal matrix. That is, in general the regression estimators are mutually correlated even when the "true" values are assumed uncorrelated.

Further,

$$\begin{aligned} E\{\mathbf{x}\hat{\mathbf{x}}'\} &= E\{\mathbf{x}\mathbf{y}'\mathbf{R}^{-1}\mathbf{F}\} \\ &= E\{\mathbf{x}\mathbf{y}'\}\mathbf{R}^{-1}\mathbf{F} \end{aligned}$$

or

$$E\{\mathbf{x}\hat{\mathbf{x}}'\} = \mathbf{F}'\mathbf{R}^{-1}\mathbf{F} \tag{5.4.10}$$

That is, in general the regression estimators are correlated with noncorresponding "true" values as well as with the corresponding "true" values.

In contrast,

$$\begin{aligned} E\{\mathbf{x}\hat{\mathbf{x}}'\} &= (\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1}\mathbf{F}'\mathbf{U}^{-2}\, E\{\mathbf{y}\mathbf{y}'\}\mathbf{U}^{-2}\mathbf{F}(\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1} \\ &= (\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1}\mathbf{F}'\mathbf{U}^{-2}\mathbf{R}\mathbf{U}^{-2}\mathbf{F}(\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1} \\ &= (\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1}\mathbf{F}'\mathbf{U}^{-2}(\mathbf{F}\mathbf{F}' + \mathbf{U}^2)\mathbf{U}^{-2}\mathbf{F}(\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1} \end{aligned}$$

from which we obtain

$$E\{\hat{\mathbf{x}}\hat{\mathbf{x}}'\} = \mathbf{I}_m + (\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1} \tag{5.4.11}$$

and similarly

$$\begin{aligned} E\{\mathbf{x}\hat{\mathbf{x}}'\} &= E\{\mathbf{x}\mathbf{y}'\mathbf{U}^{-2}\mathbf{F}(\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1}\} \\ &= \mathbf{F}'\mathbf{U}^{-2}\mathbf{F}(\mathbf{F}'\mathbf{U}^{-2}\mathbf{F})^{-1} \\ &= \mathbf{I}_m \end{aligned}$$

that is, the ML/WLSR estimators are uncorrelated with noncorresponding true factors.

It should also be noted that we can express (5.4.3) in the form

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{F}\mathbf{P} = \mathbf{U}^{-2}\mathbf{F}[\mathbf{F}'\mathbf{U}^{-2}\mathbf{F} + \mathbf{P}^{-1}]^{-1} \tag{5.4.12}$$

This is a well-known "shortcut" expression for computing the regression weights, due originally to Ledermann. It has been used in Table 5.2.1. It requires the inversion of an $m \times m$ matrix instead of an $n \times n$ matrix, thus saving arithmetic. The reader may prove the identity of the expressions in (5.4.12) by multiplying on the left by $\mathbf{R}$ in the form $\mathbf{F}\mathbf{P}\mathbf{F}' + \mathbf{U}^2$.

6 Problems of Relationship Between Factor Analyses

6.1. THE COMPARISON OF SEPARATE ANALYSES

So far we have considered factor-analytic hypotheses relating to a single-sample correlation matrix drawn from a single population. In this and Section 6.2 we consider the problem that arises when we wish to compare and contrast sets of factor-analytic results from two or more populations. In section 6.3 we consider the distinct but similar problem in repeated-measures designs where we compare factor-analytic results from the same measures repeatedly administered to the same subjects in two or more conditions. The first problem in its most general form arises when we have multivariate data from two or more samples of subjects, based on variables that might be the same or might be overlapping sets or different yet similar in what is deemed to be measured, and we wish to make comparative judgments. The comparison may be based on raw data matrices, or it may be based on our own data relative to published, possibly quite ancient, correlation matrices or published factor patterns whose origins in data have been left a total mystery.

Here we briefly consider a list of problems; then in Section 6.2 we describe a general system, due to Jöreskog, that handles a number of situations very well.

(a) Factorial Invariance

The question is, to what extent will a variable retain its *factorial description* (i.e., the list of factor loadings on the m factors that describes the variable as a mixture of factors) independently of the set of other variables in the matrix and independently of the population sample? It seems just obvious that we should not