CHAPTER 4
THE COMMON FACTOR MODEL IN THE SAMPLE

From
Exploratory Factor Analysis
Ledyard R Tucker
and
Robert C. MacCallum

CHAPTER 4
THE COMMON FACTOR MODEL IN THE SAMPLE

4.0. Introduction

In Chapter 3, we presented a mathematical representation of common factor theory in the context of a population. As was noted at the time, this was a theoretical presentation, which did not treat issues involved in sampling. Obviously, the application of factor analysis in practice requires that the model be expressed in the context of a sample, and that methods be developed for fitting the model to sample data. The purpose of the present chapter is to show how the common factor model can be represented in a sample. The mathematical framework to be developed for this purpose will serve as the basis for the treatment in later chapters of the problem of estimating the parameters of the model.

4.1. Representation of the Common Factor Model in a Sample

The representation of the common factor model in a sample will be developed by employing concepts and relationships presented in Chapter 3, which showed how the model is defined in a population. We will examine how these concepts and relationships are affected when we attempt to apply the same model to a sample drawn from the population. Given this objective, it will be useful to begin by briefly reviewing a few important equations and issues discussed in Chapter 3. A number of equations from Chapter 3 will be re-stated here for convenience and review. Considering first the expression of the model in terms of modeled attributes, recall that the population covariance matrix for the modeled attributes, $\boldsymbol{\Sigma}_{zz}$, can be defined as having the following structure:

$$\boldsymbol{\Sigma}_{zz} = \boldsymbol{\Omega}_{\beta u}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Omega}'_{\beta u} \tag{4.1}$$

In this equation, $\boldsymbol{\Omega}_{\beta u}$ is the population factor weight matrix containing submatrices $\boldsymbol{B}$, whose entries are the weights for the common factors, and $\boldsymbol{U}$, whose entries are the weights for the unique factors. The representation of $\boldsymbol{\Omega}_{\beta u}$ as a supermatrix is given by

$$\boldsymbol{\Omega}_{\beta u} = [\boldsymbol{B}, \boldsymbol{U}] \tag{4.2}$$

The matrix $\boldsymbol{\Sigma}_{xx}$ is the covariance matrix for the common and unique factors, and can be represented as a supermatrix of the following form:

$$\boldsymbol{\Sigma}_{xx} = \begin{bmatrix} \boldsymbol{\Sigma}_{\beta\beta} & \boldsymbol{\Sigma}_{\beta u} \\ \boldsymbol{\Sigma}_{u\beta} & \boldsymbol{\Sigma}_{uu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \tag{4.3}$$

As discussed in Chapter 3, given the nature of the factors and the imposed condition that they are in standardized form in the population, matrix $\boldsymbol{\Phi}$ will be a correlation matrix for the common factors, and the unique factors will be uncorrelated with each other and with the common factors. Substituting from Eqs. (4.2) and (4.3) into (4. 1) yields the oblique common factor model:

$$\boldsymbol{\Sigma}_{zz} = \boldsymbol{B\Phi B}' + \boldsymbol{U}^2 \tag{4.4}$$

The orthogonal common factor model can be represented as

$$\boldsymbol{\Sigma}_{zz} = \boldsymbol{AA}'+\boldsymbol{U}^2 \tag{4.5}$$

Chapter 3 also defined the framework for the transformation of an orthogonal solution into an oblique solution. Considering only the weight matrices for present purposes, the transformation process requires defining a population trait matrix $\boldsymbol{T}$, and the relationships between the orthogonal and oblique solutions are given as follows:

$$\boldsymbol{B} = \boldsymbol{AT}^{-1} \tag{4.6}$$

$$\boldsymbol{\Phi} = \boldsymbol{TT}' \tag{4.7}$$

Another important issue emphasized in Chapter 3 involved the distinction between modeled attributes, contained in vector $\underline{\mathbf{z}}$, and surface attributes, contained in vector $\boldsymbol{y}$. Recall that the relationship is given by

$$\boldsymbol{y} = \underline{\mathbf{z}} + \ddot{\underline{\mathbf{z}}} \tag{4.8}$$

where $\ddot{\underline{\mathbf{z}}}$ represents that part of the surface attributes that is not consistent with the common factor model. In terms of covariance matrices, the following relationship was developed in Chapter 3:

$$\boldsymbol{\Sigma}_{yy} = \boldsymbol{\Sigma}_{zz} + \boldsymbol{\Delta}_{\Sigma} \tag{4.9}$$

$$\boldsymbol{\Delta}_{\Sigma} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{zz} \tag{4.10}$$

Finally, by substitution from Eq. (4.4) into Eq. (4.9) we obtain

$$\boldsymbol{\Sigma}_{yy} = \boldsymbol{B\Phi B}' + \boldsymbol{U}^2 + \boldsymbol{\Delta}_{\Sigma} \tag{4.11}$$

This expression states the relationship between the population variances and covariances for the surface attributes, in $\boldsymbol{\Sigma}_{yy}$, and the parameters of the common factor model, in $\boldsymbol{B}$**,** $\boldsymbol{\Phi}$**,** and $\boldsymbol{U}^2$**.** As explained in Chapter 3, the entries in $\boldsymbol{\Delta}_{\Sigma}$ represent model error, in the sense that they reflect the degree to which the population covariances among the surface attributes cannot be accounted for by the common factor model. The reader also should recall the relationship of these developments to the issue of fitting the common factor model to data. If the model were being fit to $\boldsymbol{\Sigma_{yy}}$ as represented in Eq. (4.11), the objective would be to find a solution which would yield an optimal $\boldsymbol{\Delta_\Sigma}$ **.** Alternative approaches to defining what is an optimal $\boldsymbol{\Delta_\Sigma}$ would yield different solutions; i.e., different $\boldsymbol{B}$, $\boldsymbol{\Phi}$, and $\boldsymbol{U^2}$ matrices, and in turn different $\boldsymbol{\Sigma_{zz}}$ and $\boldsymbol{\Delta_\Sigma}$ matrices, obtained by Eqs. (4.4) and (4.10), respectively. Thus, it is not the case that there is a single true $\boldsymbol{\Sigma_{zz}}$ matrix, and corresponding $\boldsymbol{\Delta_\Sigma}$ matrix; rather, different $\boldsymbol{\Sigma_{zz}}$ and $\boldsymbol{\Delta_\Sigma}$ matrices would be defined for each different approach to optimizing $\boldsymbol{\Delta_\Sigma}$.

Though this is an important issue, it should be recognized that, at this point, it is stated in terms of fitting the model to a population covariance matrix. Obviously, to conduct factor analysis in practice it is necessary to represent the model in terms of a sample covariance matrix and to solve the problem of how to fit the model to such data. We will now consider the problem of expressing the model in terms of a sample covariance matrix. We wish to emphasize that this process is not as simple and straightforward as it is often taken to be. In our view, there are a number of subtle issues inherent in this problem which are often oversimplified, or completely overlooked. Of greatest importance are the distinctions between model error and sampling error, and between modeled attributes and surface attributes. We will attempt to present a framework which allows for explicit representation of these issues, along with a careful treatment of the distinction between populations and samples.

To begin, let us assume that a sample of N observations has been obtained, and that a vector $\boldsymbol{y}$ containing measures on surface attributes has been obtained from each individual. Each such vector still can be conceived of as arising from two sources, as defined in Eq. (4.8); i.e., a portion consistent with the common factor model, and a portion inconsistent with the model. This distinction between modeled attributes and surface attributes is as valid in the sample as it is in the population. Therefore, we can define a sample covariance matrix $\boldsymbol{C_{zz}}$ whose entries represent sample variances and covariances for the modeled attributes. Keep in mind that $\boldsymbol{C_{zz}}$ is not directly observable, but that it represents that portion of the sample variances and covariances of the surface attributes which can be accounted for by the common factor model. Following the form of Eq. (4.1), we can express the common factor model in terms of $\boldsymbol{C_{zz}}$ as follows:

$$\boldsymbol{C_{zz}} = \boldsymbol{\Omega_{\beta\mu} C_{xx} \Omega'_{\beta\mu}} \tag{4.12}$$

In this equation, $\boldsymbol{C_{xx}}$ represents a sample covariance matrix for the factors; i.e., this is the sample matrix corresponding to $\boldsymbol{\Sigma_{xx}}$ and would take the following form:

$$\boldsymbol{C_{xx}} = \begin{bmatrix} \boldsymbol{C_{\beta\beta}} & \boldsymbol{C_{\beta\mu}} \\ \boldsymbol{C_{\mu\beta}} & \boldsymbol{C_{\mu\mu}} \end{bmatrix} \tag{4.13}$$

A comparison of Eqs. (4.12) and (4.13) to the corresponding equations for the population, given in (4.1) and (4.3) is of central important. It is essential to recognize and understand that the weight matrix for the factors, given by $\boldsymbol{\Omega}_{\beta\mu}$, is shown as being the same in the sample as in the population. A consideration of what these factor weights represent will verify that this must be the case. Recall that the fundamental common factor model, as shown in Eq. (3.3), states that the entries in any sampled vector $\underline{\mathbf{z}}$ will be linear combinations of the entries in a vector $\underline{\boldsymbol{x}}$, which contains the measures on the factors. The weights defining these linear combinations, which are the entries in $\boldsymbol{\Omega}_{\beta\mu}$, will be the same for all such vectors. Thus, the weight matrix relating factors to modeled attributes is the same in a sample as in the full population. This relationship, however, does not imply that a common factor weight matrix obtained from an analysis of sample data will be identical to the population common factor weight matrix. In fact, as will be seen in subsequent developments in this section, several effects preclude exact determination of population common factor weights from a sample. The relationship stated here simply indicates that the sampling process per se is not the problem. The sampling process does however have an effect on the variances and covariances of the factors, both common and unique. Due to the chance characteristics of the sample, the variances and covariances of the common and unique factors in the sample will almost surely be different than the corresponding values in the population. As a result, the matrix $\boldsymbol{C}_{\boldsymbol{xx}}$ will be different from its population counterpart $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$.

A very important representation of the model in a sample can be obtained by substituting from Eqs. (4.2) and (4.13) into (4.12). This yields the following:

$$\begin{aligned}\boldsymbol{C}_{zz} &= [\boldsymbol{B},\, \boldsymbol{U}]\begin{bmatrix}\boldsymbol{C}_{\beta\beta} & \boldsymbol{C}_{\beta\mu}\\ \boldsymbol{C}_{\mu\beta} & \boldsymbol{C}_{\mu\mu}\end{bmatrix}\begin{bmatrix}\boldsymbol{B}'\\ \boldsymbol{U}'\end{bmatrix}\\ &= \boldsymbol{B}\boldsymbol{C}_{\beta\beta}\boldsymbol{B}' + \boldsymbol{B}\boldsymbol{C}_{\beta\mu}\boldsymbol{U}' + \boldsymbol{U}\boldsymbol{C}_{\mu\beta}\boldsymbol{B}' + \boldsymbol{U}\boldsymbol{C}_{\mu\mu}\boldsymbol{U}' \quad \textbf{(4.14)}\end{aligned}$$

This expression can be regarded as representing a full statement of the common factor model in the sample. However, this is not the model actually employed in practice. The model employed in practice is obtained as a result of some simplifying assumptions applied to the expression in Eq. (4.14). To see how this is accomplished, let us consider the hypothetical case in which the factor scores in the sample have some of the properties which characterize the factor scores in the population. In particular, suppose that in the sample the unique factors were uncorrelated with each other and with the common factors. This would result in all entries in matrices $\boldsymbol{C}_{\beta\mu}$ and $\boldsymbol{C}_{\mu\beta}$ being zero, and matrix $\boldsymbol{C}_{\mu\mu}$ being diagonal. Furthermore, let us suppose that the unique factors are scaled so that they have unit variances, thus yielding a matrix $\boldsymbol{C}_{\mu\mu}$ which is an identity matrix. Under these conditions, the model in Eq. (4.14) would become

$$\boldsymbol{C}_{zz} = \boldsymbol{B}\boldsymbol{C}_{\beta\beta}\boldsymbol{B}' + \boldsymbol{U}^2 \tag{4.15}$$

If this model were a valid representation of the true structure of $\boldsymbol{C}_{zz}$, then it would imply that the sampling process has no effect on the common factor weights or unique variances, and only affects the common factor covariances. However, it is critical to recognize that the hypothetical conditions which yielded this equation will almost certainly not hold in practice. That is, the process of sampling will almost surely give rise to a sample in which correlations of the unique factors with each other and with the common factors are not exactly zero. As a result, matrices $\boldsymbol{C}_{\beta\mu}$ and $\boldsymbol{C}_{\mu\beta}$ will not actually be zero, and matrix $\boldsymbol{C}_{\mu\mu}$ will not actually be diagonal. This in turn implies that matrix $\boldsymbol{C}_{zz}$ cannot be exactly represented by a model which is based on these assumptions; i.e., the model in Eq. (4.15). This context is helpful in moving toward an understanding of how factor analysis is conducted in practice, and of what sources contribute to a lack of perfect fit of the model to real data. In effect, and this is a very important point, the representation of the common factor model in the sample is defined in practice as in Eq. (4.15). The implication is that in practice we treat the sampling issue as if unique factors were uncorrelated with each other and with the common factors in the sample. Since this will not in fact be the case in the real world, the matrix $\boldsymbol{C}_{zz}$ cannot actually be fit exactly by the model in Eq. (4.15). To represent this mathematically, it is useful to define a matrix $\boldsymbol{\Delta}_z$ whose entries represent a lack of fit in that model. The model could then more appropriately be written as

$$\boldsymbol{C}_{zz} = \boldsymbol{B}\boldsymbol{C}_{\beta\beta}\boldsymbol{B}' + \boldsymbol{U}^2 + \boldsymbol{\Delta}_z \tag{4.16}$$

The entries in $\boldsymbol{\Delta}_z$ can be viewed as reflecting a primary source of sampling error. That is, if the characteristics of the sample were identical to those of the population, then all entries in $\boldsymbol{\Delta}_z$ would be zero, and the model in Eq. (4.15) would be valid. However, the phenomenon of sampling error gives rise to samples with properties not identical to the population. In the present context, this causes the model in Eq. (4.15) to be inexact, with the degree of error reflected in the entries in $\boldsymbol{\Delta}_z$ in Eq. (4.16). In other words, the model in Eq. (4.15) is based on simplifying assumptions which generally do not hold in practice, and this fact gives rise to a primary source of sampling error in the model, as represented in Eq. (4.16). If the assumptions did hold exactly in a sample, then there would be no sampling error of this type; in that case, all entries in $\boldsymbol{\Delta}_z$ in Eq. (4.16) would be zero and the simplified model in Eq. (4.15) would be exactly correct. This framework thus shows explicitly the manner in which this source of sampling error impacts on the common factor model. Other potential source of sampling error will be considered later in this chapter and in Chapter 5.

A remaining aspect of this representation of the model which requires attention is the issue of the standardization of the factors. As expressed in Eq. (4.16) the model is stated in terms of unstandardized common factors; that is, though we have considered the common factors to be standardized in the population, they almost surely will not be such in the sample. This can be easily alleviated by converting the covariance matrix $\boldsymbol{C}_{\beta\beta}$ in Eq. (4.16) into a correlation matrix. In fact, this standardization of the common factors in the sample is the most common approach taken to resolving a critical problem arising in the process of fitting the common factor model to real data. This problem, called the identification problem, involves the fact that when the model is expressed as in Eq. (4.16), the scales of the common factors are completely arbitrary. In order to estimate the parameters of the model, it is necessary to establish a scale, or unit of measurement, for the factors in the sample. The simplest and most common way to achieve this is to define the common factors as being standardized in the sample. This entire issue will be discussed in more detail in Chapters 8 and 9. For present purposes, let us define a mathematical representation of the model where the common factors are standardized in the sample.

Let us first define a diagonal matrix $[\boldsymbol{C_d}]_{\beta\beta}$ containing the variances of the common factors in the sample; that is,

$$[\boldsymbol{C_d}]_{\beta\beta} = \boldsymbol{Diag}(\boldsymbol{C}_{\beta\beta}) \tag{4.17}$$

We can then define a matrix $\widetilde{\boldsymbol{C}}_{\beta\beta}$ which is the correlation matrix for the common factor in the sample as follows:

$$\widetilde{\boldsymbol{C}}_{\beta\beta} = [\boldsymbol{C_d}]_{\beta\beta}^{-\frac{1}{2}} \boldsymbol{C}_{\beta\beta} [\boldsymbol{C_d}]_{\beta\beta}^{-\frac{1}{2}} \tag{4.18}$$

Let us next define a matrix $\widetilde{\boldsymbol{B}}$ as follows:

$$\widetilde{\boldsymbol{B}} = \boldsymbol{B}[\boldsymbol{C_d}]_{\beta\beta}^{\frac{1}{2}} \tag{4.19}$$

Note that $\widetilde{\boldsymbol{B}}$ contains population factor weights which have been re-scaled so as to correspond to common factors standardized in the sample. That is, the columns of $\boldsymbol{B}$, the matrix of population common factor weights, are each multiplied by the sample standard deviation of the corresponding common factor to produce the weights in $\widetilde{\boldsymbol{B}}$. Given these definitions, it can be seen that

$$\begin{aligned} \widetilde{\boldsymbol{B}}\, \widetilde{\boldsymbol{C}}_{\beta\beta} \widetilde{\boldsymbol{B}}\,' &= \boldsymbol{B}[\boldsymbol{C_d}]_{\beta\beta}^{\frac{1}{2}} [\boldsymbol{C_d}]_{\beta\beta}^{-\frac{1}{2}} \boldsymbol{C}_{\beta\beta} [\boldsymbol{C_d}]_{\beta\beta}^{-\frac{1}{2}} [\boldsymbol{C_d}]_{\beta\beta}^{\frac{1}{2}} \boldsymbol{B}' \\ &= \boldsymbol{B}\boldsymbol{C}_{\beta\beta}\boldsymbol{B}\,' \end{aligned} \tag{4.20}$$

This relationship is important because it allows us to re-write the model in Eq. (4.16) in terms of common factors standardized in the sample. That is, by substituting from Eq. (4.20) into Eq. (4.16) we obtain

$$\boldsymbol{C}_{zz} = \widetilde{\boldsymbol{B}}\, \widetilde{\boldsymbol{C}}_{\beta\beta} \widetilde{\boldsymbol{B}}\,' + \boldsymbol{U^2} + \boldsymbol{\Delta}_z \tag{4.21}$$

Note that $\boldsymbol{U^2}$ and $\boldsymbol{\Delta}_z$ are the same in Eqs. (4.21) and (4.16); the distinction between the two expressions is that the common factors in the latter are defined as standardized in the sample.

Let us now consider the final step in expressing the model in terms of observable data; i.e., the step from modeled attributes to surface attributes. The preceding developments were set in the context of modeled attributes, with the common factor model being stated in terms of the sample covariance matrix for the modeled attributes, $\boldsymbol{C_{zz}}$. As noted above, the objective is to express the model in terms of the sample covariance matrix for the surface attributes, which will be designated $\boldsymbol{C_{yy}}$. We must first consider the relationship between $\boldsymbol{C_{zz}}$ and $\boldsymbol{C_{yy}}$. This relationship follows the same pattern as the relationship between $\boldsymbol{\Sigma_{zz}}$ and $\boldsymbol{\Sigma_{yy}}$, which was derived in
Chapter 3 and is shown in Eq. (3.88). By following the same procedure, we can state the following relationship:

$$\boldsymbol{C_{yy}} = \boldsymbol{C_{zz}} + \boldsymbol{C_{z\ddot{z}}} + \boldsymbol{C_{\ddot{z}z}} + \boldsymbol{C_{\ddot{z}\ddot{z}}} \tag{4.22}$$

Following the developments presented in Chapter 3 for the population, let us define a matrix $\boldsymbol{\Delta_c}$ as follows:

$$\boldsymbol{\Delta_c} = \boldsymbol{C_{z\ddot{z}}} + \boldsymbol{C_{\ddot{z}z}} + \boldsymbol{C_{\ddot{z}\ddot{z}}} \tag{4.23}$$

Substituting from Eq. (4.23) into Eq. (4.22), we can write

$$\boldsymbol{C_{yy}} = \boldsymbol{C_{zz}} + \boldsymbol{\Delta_c} \tag{4.24}$$

The matrix $\boldsymbol{\Delta_c}$ could alternatively be defined as

$$\boldsymbol{\Delta_c} = \boldsymbol{C_{yy}} - \boldsymbol{C_{zz}} \tag{4.25}$$

The elements in $\boldsymbol{\Delta_c}$ can be viewed as reflecting a lack of correspondence between $\boldsymbol{C_{yy}}$ and $\boldsymbol{C_{zz}}$. It is very interesting to examine the source of this lack of correspondence. The entries in $\boldsymbol{\Delta_c}$ can be thought of as arising from model error. That is, since the matrix $\boldsymbol{C_{zz}}$ represents that portion of the sample variances and covariances of the surface attributes that can be accounted for by the model, the entries in $\boldsymbol{\Delta_c}$ arise from lack of fit of the model to the sample data. Thus, matrix $\boldsymbol{\Delta_c}$, representing model error in the sample, is analogous to matrix $\boldsymbol{\Delta_\Sigma}$, defined in Eq. (4.10), which represents model error in the population.

It is now possible to combine preceding developments to achieve a unified view of model error and sampling error, and to produce a representation of the common factor model in terms of a sample covariance matrix for surface attributes. By substituting from Eq. (4.21) into Eq. (4.24), we obtain

$$\boldsymbol{C_{yy}} = (\widetilde{\boldsymbol{B}}\,\widetilde{\boldsymbol{C}}_{\beta\beta}\widetilde{\boldsymbol{B}}' + \boldsymbol{U}^2 + \boldsymbol{\Delta}_z) + \boldsymbol{\Delta_c} \tag{4.26}$$

If we define a matrix $\boldsymbol{\Delta_y}$ according to

$$\boldsymbol{\Delta}_y = \boldsymbol{\Delta}_z + \boldsymbol{\Delta}_c \tag{4.27}$$

and substitute from Eq. (4.27) into Eq. (4.26), we obtain

$$\boldsymbol{C}_{yy} = \widetilde{\boldsymbol{B}}\,\tilde{\boldsymbol{C}}_{\beta\beta}\widetilde{\boldsymbol{B}}' + \boldsymbol{U}^2 + \boldsymbol{\Delta}_y \tag{4.28}$$

Given the interpretation of $\boldsymbol{\Delta}_z$, defined in conjunction with Eq. (4.16), as reflecting the impact of sampling error, and of $\boldsymbol{\Delta}_c$ as reflecting the impact of model error in the sample, it can be seen that the entries in $\boldsymbol{\Delta}_y$ reflect both sources of error. That is, the lack of fit of the model arises from two sources: (a) the fact that the simplifying assumptions made in conjunction with Eq. (4.14) will not hold exactly in practice, thus giving rise to sampling error; and (b) the fact that the common factor model is not expected to fit exactly the covariances of the surface attributes, thus giving rise to model error. It is not possible in practice to actually separate sampling error from model error. This would require knowing true values of population parameters. However, the developments just presented do serve to separate these two sources of error in theoretical terms, and to provide an explanation of their separate and combined impact on the representation of the model in a sample.

Eq. (4.28) is of central importance in understanding and solving the problem of fitting the common factor model to sample data. An objective of Chapters 3 and 4 has been to reach this point of representing the structure of a sample covariance matrix for surface attributes in terms of the parameters of the common factor model. This is achieved by Eq. (4.28). The equation expresses the sample covariance matrix for the surface attributes as a function of common factor weights, common factor intercorrelations, unique variances, and a "lack of fit" term. Eq. (4.28) is the sample equivalent of Eq. (4.11), which expresses the population covariance matrix for the surface attributes as a function of the model parameters. The representation of the model in the sample is the basis for developing methods for fitting the model to sample covariance matrices. In basic terms, the problem is as follows: given matrix $\boldsymbol{C}_{yy}$, we wish to obtain coefficient matrices $\widetilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}_{\beta\beta}$, and $\boldsymbol{U}^2$ so that the entries in $\boldsymbol{\Delta}_y$ are in some sense made optimally small. There is a close relation between the sense in which $\boldsymbol{\Delta}_\Sigma$ is optimal and the sense in which $\boldsymbol{\Delta}_y$ is optimal. Remember that the matrices $\boldsymbol{B}, \boldsymbol{\Phi}$, and $\boldsymbol{U}^2$ depend on the way in which $\boldsymbol{\Delta}_\Sigma$ is optimal (see Eq. (4.11) and ensuing discussion). A consequence is that the population parameters which are estimated by matrices $\widetilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}_{\beta\beta}$, and $\boldsymbol{U}^2$ determined from $\boldsymbol{C}_{yy}$ depend on the sense in which $\boldsymbol{\Delta}_\Sigma$ and $\boldsymbol{\Delta}_y$ are optimal. In the following discussion, the sense in which optimality is defined will be taken to be fixed.

Several important aspects of this problem must be carefully considered. The first involves the nature of the parameters to be estimated. Note that entries in $\widetilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}_{\beta\beta}$, and $\boldsymbol{U}^2$ are not all pure population parameters unaffected by characteristics of the sample. Specifically, the entries in $\widetilde{\boldsymbol{B}}$ are population common factor weights which have been rescaled to represent common factors standardized in the sample. An obtained estimate of $\widetilde{\boldsymbol{B}}$ can be viewed as an estimate of $\boldsymbol{B}$, with some error of estimation arising from the fact that the factors are standardized in the sample rather than the population. In a similar fashion, the entries in $\tilde{\boldsymbol{C}}_{\boldsymbol{\beta\beta}}$ are actually sample correlations among the common factors. However, these sample values cannot be observed or determined exactly because they are defined in terms of the population common factor weights in $\boldsymbol{B}$. An obtained estimate of $\tilde{\boldsymbol{C}}_{\boldsymbol{\beta\beta}}$ can be viewed as an estimate of $\boldsymbol{\Phi}$; again, some error of estimation of $\boldsymbol{\Phi}$ arises from the fact that the factors are standardized in the sample rather than the population.

This view helps to achieve an understanding of what is being estimated when the model given in Eq. (4.28) is fit to sample data. Though the parameters in that model are represented as $\widetilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}_{\boldsymbol{\beta\beta}}$, and $\boldsymbol{U}^{\mathbf{2}}$, that representation incorporates the fact that the common factors are standardized in the sample. That standardization will most likely be at least slightly different from standardization in the population and is thus the source of some degree of sampling error. This sampling error occurs in addition to that defined by $\boldsymbol{\Delta}_{\boldsymbol{z}}$ in Eq. (4.16) and incorporated into $\boldsymbol{\Delta}_{\boldsymbol{y}}$ in Eq. (4.27). Thus, when the model in Eq. (4.28) is fit to sample data, an optimal solution for the values in $\widetilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}_{\boldsymbol{\beta\beta}}$, and $\boldsymbol{U}^{\mathbf{2}}$ can be seen to provide estimates of the parameters in $\boldsymbol{B}$, $\boldsymbol{\Phi}$, and $\boldsymbol{U}^{\mathbf{2}}$. Error in those estimates arises from the sampling error and model error included in $\boldsymbol{\Delta}_{\boldsymbol{y}}$ as well as from the standardization of the factors in the sample as defined by Eq. (4.18) and (4.19).

A second important aspect of this problem involves the fact that, as has been noted in other similar contexts previously, there are alternative possible definitions of what would be an optimal $\boldsymbol{\Delta}_{\boldsymbol{y}}$. Each alternative definition would yield an alternative method for estimating the model parameters, and, in turn, a different set of estimates of those parameters. While alternative methods for defining optimal fit and estimating parameters will be discussed in Chapter 7, it is important at the present time to understand that parameter estimates can be obtained in different ways, and to recognize how this fact impacts on the developments presented in this section. Let us define matrices of parameter estimates obtained by any desired method. Matrix $\boldsymbol{B}$ will be an $n \times r$ matrix of estimated common factor weights, with element b_{jk} representing the estimated weight for factor *k* on attribute *j*. Matrix $\boldsymbol{R}_{\boldsymbol{\beta\beta}}$ will be an $r \times r$ matrix of estimated common factor intercorrelations, with entry r_{kl}, representing the estimated correlation of factors *k* and *l*. Finally, matrix $\boldsymbol{U}^{\mathbf{2}}$ will be an $n \times n$ diagonal matrix, with diagonal entry u_{jj} representing the estimated unique variance for attribute *j*. For any given solution, the resulting matrices $\boldsymbol{B}$, $\boldsymbol{R}_{\boldsymbol{\beta\beta}}$, and $\boldsymbol{U}^{\mathbf{2}}$

can be employed to obtain a covariance matrix for the modeled attributes as represented by that solution. Let such a matrix be designated $\boldsymbol{C}_{zz}^{+}$. This matrix can be obtained as follows:

$$\boldsymbol{C}_{zz}^{+} = \boldsymbol{B}\boldsymbol{R}_{\beta\beta}\boldsymbol{B}' + \boldsymbol{U}^{2} \tag{4.29}$$

It is important to understand the distinction between $\boldsymbol{C}_{zz}$ and $\boldsymbol{C}_{zz}^{+}$**.** The matrix $\boldsymbol{C}_{zz}$ is the modeled attribute covariance matrix which is defined by the full common factor model in the sample, as given in Eq. (4.14) or (4.16). Recall that simplifying assumptions are made regarding that full model, resulting in the simplified model given in Eq. (4.15). Matrix $\boldsymbol{C}_{zz}^{+}$ is thus a modeled attribute covariance matrix constructed from a solution by substituting parameter estimates into Eq. (4.15). The matrix $\boldsymbol{C}_{zz}$ cannot be exactly determined in practice because it is defined in terms of true parameter values. However, matrix $\boldsymbol{C}_{zz}^{+}$ is a very useful matrix which can be obtained from any common factor solution. It represents the covariance matrix for the modeled attributes as represented <u>by that solution</u>. The greater the correspondence between $\boldsymbol{C}_{zz}^{+}$ and $\boldsymbol{C}_{yy}$, the more closely the model is found to fit the sample data. An important point is that there is no single "true" $\boldsymbol{C}_{zz}^{+}$ matrix. Rather, there is a different $\boldsymbol{C}_{zz}^{+}$ for each different method of fitting the model to $\boldsymbol{C}_{yy}$. That is, different definitions of an optimal solution to the model in Eq. (4.28) will yield different $\boldsymbol{B}$, $\boldsymbol{R}_{\beta\beta}$, and $\boldsymbol{U}^{2}$ matrices, and, in turn, different $\boldsymbol{C}_{zz}^{+}$ matrices. Recall that the same phenomenon was discussed earlier in this section in the context of the population. The present development merely shows how this phenomenon occurs in the sample.

When a factor solution is obtained and the corresponding $\boldsymbol{C}_{zz}^{+}$ matrix is calculated, it is then possible to determine a matrix representing the lack of fit of that solution to sample data. This matrix will be designated as $\boldsymbol{\Delta}_{y}^{+}$ will be defined as follows:

$$\boldsymbol{\Delta}_{y}^{+} = \boldsymbol{C}_{yy} - \boldsymbol{C}_{zz}^{+} \tag{4.30}$$

This matrix represents the optimal solution for $\boldsymbol{\Delta}_{y}$ obtained as a result of the application of some specific method for estimating the parameters of the model. The entries in $\boldsymbol{\Delta}_{y}^{+}$ arise from a combination of model error, which is affected by the method employed to fit the model to the data, and sampling error.

A final point about the parameter estimation problem is that some additional complexities will arise when model fitting procedures are considered in Chapters 8 and 9. Though the model represented by Eq. (4.28) provides a useful framework for the model fitting problem, it will be seen that this framework requires slight modification for some methods. Since this modification will involve a rescaling of the parameters, it is not the case that all model fitting procedures are providing estimates of the exactly the same parameters. This complication will be discussed in detail in Chapters 8 and 9.

We wish to close this section by noting two important points which will be treated at length in subsequent chapters. First, though the model given in Eq. (4.28) is expressed in terms of oblique common factors, most fitting procedures obtain solutions in which the common factors are orthogonal. This will be discusses in Chapters 8 and 9. Second, these obtained orthogonal common factors then can be transformed into oblique common factors. This transformation procedure was discussed in detail in Chapter 3 in the context of the population. For present purposes, suffice it to say that this freedom to transform or "rotate" factors exists in obtained sample solutions as well. The mathematical framework for this process, along with methods for seeking meaningful transformed factors, will be presented in Chapters 10 and 11.

4.3. The Common Factor Model for a Sample Correlation Matrix

The developments in the previous section were carried out in the context of a sample covariance matrix. The model given by Eq. (4.28) defines the factorial structure of such a matrix. An issue which has been raised previously in this book involves the distinction between factor analyzing covariance vs. correlation matrices. This issue was treated in Chapter 3 in the context of the population, and will be considered here in the context of the sample. Following the developments in Section 3.6, let us begin by representing the conversion of a sample covariance matrix, $\boldsymbol{C_{yy}}$, into a sample correlation matrix, $\boldsymbol{R_{yy}}$**.** This is achieved by first defining a diagonal matrix $\boldsymbol{[C_d]_{yy}}$ containing the sample variance of the surface attributes. That is,

$$\boldsymbol{[C_d]_{yy} = Diag(C_{yy})} \tag{4.31}$$

We then can define a sample correlation matrix, $\boldsymbol{R_{yy}}$, for the surface attributes as follows:

$$\boldsymbol{R_{yy}} = [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \boldsymbol{C_{yy}} [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \tag{4.32}$$

Effects of this standardization of the surface attributes on the common factor model in the sample can be seen by substituting from Eq. (4.28) into Eq. (4.32). This yields:

$$\begin{aligned} \boldsymbol{R_{yy}} &= [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} (\widetilde{\boldsymbol{B}}\,\widetilde{\boldsymbol{C}}_{\boldsymbol{\beta\beta}} \widetilde{\boldsymbol{B}}\,' + \boldsymbol{U^2} + \boldsymbol{\Delta}_{\boldsymbol{y}}) [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \\ &= [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \widetilde{\boldsymbol{B}}\,\widetilde{\boldsymbol{C}}_{\boldsymbol{\beta\beta}} \widetilde{\boldsymbol{B}}\,' [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} + [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \boldsymbol{U^2} [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \\ &\quad + [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \boldsymbol{\Delta}_{\boldsymbol{y}} [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \end{aligned} \tag{4.33}$$

To simplify this expression, we define the following matrices:

$$\widetilde{\widetilde{\boldsymbol{B}}} = [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \widetilde{\boldsymbol{B}} \tag{4.34}$$

$$\widetilde{\widetilde{\boldsymbol{U}^{\boldsymbol{2}}}} = [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \boldsymbol{U^2} [\boldsymbol{C_d}]_{\boldsymbol{yy}}^{-\frac{1}{2}} \tag{4.35}$$

$$\widetilde{\overline{\boldsymbol{\Delta}}}_y = [\boldsymbol{C}_d]_{yy}^{-\frac{1}{2}}\boldsymbol{\Delta}_y[\boldsymbol{C}_d]_{yy}^{-\frac{1}{2}} \tag{4.36}$$

These matrices contain common factor weights, unique variances, and lack of fit terms, respectively, which have been rescaled as a result of the standardization of the attributes in the sample. Substituting from these equations into Eq. (4.33) yields the following model:

$$\boldsymbol{R}_{yy} = \widetilde{\overline{\boldsymbol{B}}}\,\tilde{\boldsymbol{C}}_{\beta\beta}\widetilde{\overline{\boldsymbol{B}}}\,' + \widetilde{\overline{\boldsymbol{U}^2}} + \widetilde{\overline{\boldsymbol{\Delta}}}_y \tag{4.37}$$

This equation expresses the common factor model in terms of a sample correlation matrix for the surface attributes. It has the same general form as Eq. (4.28), which expressed the model in terms of a sample covariance matrix, but some of the terms have been rescaled. Another important comparison is to note that Eq. (4.37) has the same form as Eq. (3.98); the latter expressed the model in terms of a population correlation matrix for the surface attributes. Thus, Eq. (4.37) is the sample equation analogous to the population equation given in Eq. (3.98). When the model in Eq. (4.37) is fit to a sample correlation matrix, $\boldsymbol{R}_{yy}$, we would obtain a solution consisting of common factor weights, common factor intercorrelations, and unique variances. To represent necessary distinctions between these values and those which would be obtained from analysis of a covariance matrix, as defined in the previous section, let us define the following matrices: $\widetilde{\overline{\boldsymbol{B}}}$ contains sample common factor weights obtained from analysis of the correlation matrix; $\boldsymbol{R}_{\beta\beta}$ contains corresponding sample common factor intercorrelations; and $\widetilde{\overline{\boldsymbol{U}^2}}$ contains corresponding sample unique variances. The fact that no new notation is used for $\boldsymbol{R}_{\beta\beta}$ reflects the fact that this matrix is not affected by the analysis of a correlation matrix rather than a covariance matrix; this can be seen by comparing Eqs. (4.28) and (4.37). The sample values obtained in $\widetilde{\overline{\boldsymbol{B}}}$, $\boldsymbol{R}_{\beta\beta}$**,** and $\widetilde{\overline{\boldsymbol{U}^2}}$ can be viewed as estimates of the parameters in $\boldsymbol{B}^*$, $\boldsymbol{\Phi}$, and $\boldsymbol{U}^{*^2}$ represented in Eq. (3.98). That is, the population parameters representing the factorial structure of a population correlation matrix are estimated by obtaining a factorial solution for a sample correlation matrix.

An interesting aspect of this issue can be seen by comparing Eqs. (3.95)-(3.96) to Eqs. (4.34)-(4.35). Note that the former represent a rescaling of population parameters by population standard deviations of the attributes, while the latter represent rescaling of those same parameters by sample standard deviations of the attributes. This distinction reveals a source of sampling error which influences the accuracy of the sample values in $\widetilde{\overline{\boldsymbol{B}}}$ and $\widetilde{\overline{\boldsymbol{U}^2}}$ as estimates of the parameters in $\boldsymbol{B}^*$ and $\boldsymbol{U}^{*^2}$. This sampling error is present in addition to that contained in $\widetilde{\overline{\boldsymbol{\Delta}}}_y$**.** A more detailed discussion and demonstration of this phenomenon will be presented in Chapter 5.

To complete this presentation of the common factor model for a sample correlation matrix, note that we can follow the development given in Eq. (4.29) by defining a "reconstructed" correlation matrix , $\boldsymbol{R}_{zz}^{+}$, as follows:

$$\boldsymbol{R}_{zz}^{+} = \widetilde{\overline{\boldsymbol{B}}}\, \boldsymbol{R}_{\beta\beta} \widetilde{\overline{\boldsymbol{B}}}\,' + \widetilde{\overline{\boldsymbol{U}}}^{2} \tag{4.38}$$

This matrix is a correlation matrix for the modeled attributes constructed from the obtained factor solution. It provides the correlations among the modeled attributes as represented by that solution. We can also obtain a matrix representing the lack of fit of that solution to the sample data. This matrix will be designated $\widetilde{\overline{\boldsymbol{\Delta}}}_{y}^{+}$ and will be defines as follows:

$$\widetilde{\overline{\boldsymbol{\Delta}}}_{y}^{+} = \boldsymbol{R}_{yy} - \boldsymbol{R}_{zz}^{+} \tag{4.39}$$

This matrix represents the optimal solution for $\widetilde{\overline{\boldsymbol{\Delta}}}_{y}$ obtained as a result of the application of some specific method for estimating the parameters of the model in Eq. (4.37). As noted in other similar contexts, it should be recognized that alternative fitting methods would yield different solutions for $\boldsymbol{B}$, $\boldsymbol{R}_{\beta\beta}$, and $\boldsymbol{U}^{2}$, and thus different $\boldsymbol{R}_{zz}^{+}$ and $\widetilde{\overline{\boldsymbol{\Delta}}}_{y}^{+}$ matrices.

The important general points to recognize from the developments in this section are that (a) the common factor model for a sample correlation matrix is a special case of the model for a sample covariance matrix, and (b) the standardization of the attributes in the sample introduces an additional source of sampling error to be discussed further in Chapter 5.

4.4. Sources of Error in Fitting the Common Factor Model to Sample Data

Special attention should be given to the fact that developments in this chapter have served to identify a number of sources of error which affect solutions obtained when the common factor model is fit to sample data. These effects cause parameter estimates to be in error and unstable to some degree, and also give rise to lack of fit of the model to the data. A primary source of such error has been designated as model error. The occurrence of model error simply reflects the fact that the common factor model will generally not precisely account for all variance and covariance of the surface attributes. This may be due to such phenomena as nonlinear relations of factors to attributes, the presence of a large number of minor factors not represented in the model, etc. Algebraically, model error is represented in the population by matrix $\boldsymbol{\Delta}_{\Sigma}$ in Eq. (4.11), and in the sample by matrix $\boldsymbol{\Delta}_{c}$ in Eq. (4.26).

Another general source of error is sampling error, which includes phenomena whereby chance characteristics of a sample influence parameter estimates and the fit of the model to the data. Developments in this chapter have revealed that there exist several separate sources of sampling error. A primary source arises from the fact that the assumption that unique factors are uncorrelated with each other and with common factors will generally not hold exactly in a sample. This gives rise to the lack of fit represented in matrix $\boldsymbol{\Delta_z}$, which was discussed in conjunction with Eqs. (4.15) and (4.16). This source of sampling error would affect estimates of all parameters of the model. A second type of sampling error arises from the common procedure of standardizing the common factors in the sample. Since this standardization will tend to be at least slightly different from that in the population, this gives rise to some error in estimation of the parameters in $\boldsymbol{B}$ and $\boldsymbol{\Phi}$. This was discussed in conjunction with Eq. (4.28). An important aspect of this source of sampling error is that, as discussed previously in this chapter, standardization of the factors in the sample resolves the identification problem in parameter estimation. Since this problem will be dealt with in order to fit the model to data, some sampling error will arise no matter how it is resolved. A third source of sampling error arises when the attributes are standardized in the sample. This occurs when a sample correlation matrix rather than a sample covariance matrix is analyzed. It was shown in the previous section that this standardization introduces a source of error which would affect estimates of common factor weights and unique variances.

In practice these various sources of error cannot be separated. However, we feel it is important to have a theoretical framework in which their separate effects can be represented and understood. Some of these effects will be demonstrated and described further in Chapter 5, and implications for practice will be discussed in Chapter 6.

4.5. Illustration of Fitting the Common Factor Model to Sample Data

Rather than present a completely new illustration of the developments described in this chapter, we will briefly describe how an illustration presented in a previous chapter can be viewed in the present context. In Chapter 1 we presented results of factor analysis of data drawn from a study by Thurstone and Thurstone (1941). The data which were factor analyzed consisted of correlations among nine mental tests, based on measures obtained from a sample of 710 students. The correlation matrix shown in Table 1.1 corresponds to matrix $\boldsymbol{R_{yy}}$. The model represented by Eq. (4.37) was fit to this matrix (by methods described in Chapter 7), and the resulting solution is shown in Table 1.2. The factor weights in Table 1.2 correspond to matrix $\widetilde{\overline{\boldsymbol{B}}}$ and the factor intercorrelations correspond to matrix $\boldsymbol{R}_{\beta\beta}$. The communalities in Table 1.2 were obtained by subtracting the unique variances in $\widetilde{\overline{\boldsymbol{U}^2}}$ from unity. We could define these communalities as entries in a diagonal matrix $\widetilde{\overline{\boldsymbol{H}^2}}$, where

$$\widetilde{\overline{\boldsymbol{H}^2}} = \boldsymbol{I} - \widetilde{\overline{\boldsymbol{U}^2}} \tag{4.40}$$

According to developments in the previous section, values given in Table 1.2 represent estimates of the parameters in $\boldsymbol{B}^*$, $\boldsymbol{\Phi}$, and $\boldsymbol{H}^{*^2}$, respectively. From this solution, one could compute a correlation matrix for the modeled attributes according to Eq. (4.38) and a matrix of lack of fit terms according to Eq. (4.39). We leave this as an exercise for the reader.

The reader should keep in mind a number of influence which affect the solution obtained in this illustration. If a covariance matrix rather than a correlation matrix had been analyzed, the effects of standardization of the attributes would not have been present and the illustration would have fit the framework described in Section 4.2 rather than that in Section 4.3. The effect of this would have been a rescaling of the factor weights and unique variances, as indicated by Eqs. (4.34) and (4.35). In addition, if a different method had been used to fit the model to the data, a slightly different solution would have been obtained. Further consideration of this issue, along with additional illustrations, will be presented in Chapters 8 and 9.

4.6. Conclusion

This completes the development of the representation of the common factor model in the sample. We wish to re-emphasize that the framework presented here serves to provide an explicit treatment of a number of important issues inherent in factor analytic theory: (a) the distinction between the population and the sample; (b) the distinction between surface attributes and modeled attributes; and (c) the distinction between model error and various types of sampling error. The last two issues in particular are almost totally ignored in most presentations of factor analysis. We believe, however, that a complete theoretical understanding of the model requires that these aspects be considered very explicitly. The framework developed here will be employed in Chapters 8 and 9 to resolve the problem of fitting the model to a sample covariance matrix. Prior to that, however, some important issues involved in selecting observations and attributes in a factor analysis study will be considered in Chapter 6.