CHAPTER 3

CONSTRUCTING DATA WITH KNOWN UNDERLYING FACTOR STRUCTURE

3.1 INTRODUCTION

The main purpose of this study is to compare the performance of the four methods of analysis, presented in Chapter 2, under various circumstances. All the data analyzed in this study come from simulation. That is, data are constructed by starting from some "true" model describing features of a population. From this population samples are drawn to which noise is added. In this chapter the method of data simulation is presented. In order to quantify the performance of the methods, their respective solutions will be compared with the underlying features of the simulated samples which, as said, together constitute the true model. These features are the pattern matrix, the structure matrix, the weights matrix and the factor correlation matrix. When talking about these underlying matrices, they will be given the adjective "true", because they come from the "true" model. For the different kinds of data to be simulated in this study, the true pattern, true structure, true weights and true factor correlation matrix were demanded. For the true weights matrix this appeared to be an impossibility, as will be explained in Section 3.2.2.

3.1.1 Terminology

The following definitions will be maintained throughout the present chapter:

n_k = a variable indicating sample size in group k, k=1,...,p,

$\mathbf{Y}_k$ = an $n_k \times m$ matrix of raw scores of n_k individuals on m variables,

$\mathbf{Z}_k$ = standardized version of $\mathbf{Y}_k$,

$\mathbf{R}_k$ = the associated $m \times m$ correlation matrix,

$\mathbf{F}_k$ = $\mathbf{Z}_k\mathbf{W}_k$ = an $n_k \times q$ matrix of factor scores,

$\mathbf{\Phi}_k$ = the $q \times q$ matrix of correlations between the factors,

$\mathbf{P}_k$ = the $m \times q$ pattern matrix, containing the loadings (linear regression weights) for (re)constructing the variables from the factors,

$\mathbf{S}_k$ = $\mathbf{P}_k\mathbf{\Phi}_k$ = the $m \times q$ structure matrix, containing the correlations between the variables and the factors,

$\mathbf{W}_k$ = an $m \times q$ component weights matrix, such that $\mathbf{F}_k = \mathbf{Z}_k\mathbf{W}_k$,

$\mathbf{E}_k$ = an $n_k \times m$ random sample matrix containing n_k realizations of a stochastic multivariate variable $\mathbf{e}_k$ (of order $m \times 1$) $\sim$ $N(0,\mathbf{D}^2_{ek})$, where

$\mathbf{D}^2_{ek}$ = an $m \times m$ diagonal matrix with in the j^{th} position on the diagonal the variance (ε^2_{jk}) of the normal distribution from which the random sample for variable j is drawn,

t = subscript indicating true population matrices (i.e. $\mathbf{P}_t$, $\mathbf{S}_t$, $\mathbf{W}_t$ and $\mathbf{\Phi}_t$).

3.2 DESCRIPTION OF THE DATA SIMULATION

For the data simulation procedure, the true pattern matrices $\mathbf{P}_{tk}$, $k=1,...,p$, and the true factor correlation matrices $\mathbf{\Phi}_{tk}$, $k=1,...,p$, have to be prespecified. Also required is a diagonal matrix $\mathbf{D}^2_{ek}$, with on the j^{th} position of the diagonal the value ε^2_{jk} $(j=1,...,m)$, which stands for the prespecified variance of the normal distribution from which the error scores for variable j are drawn. Each simulated sample k consists of a prespecified number of n_k realizations of the m-dimensional variable

$$\mathbf{y}_k = \mathbf{P}_{tk}\mathbf{f}_k + \mathbf{e}_k, \tag{3.1}$$

where $\mathbf{f}_k$ is a q-dimensional random variable with a multivariate normal distribution $N(0,\mathbf{\Phi}_{tk})$ and $\mathbf{e}_k$ is an m-dimensional random (measurement error) variable with a multivariate normal distribution $N(0,\mathbf{D}^2_{ek})$.

Each sample is generated as follows. Firstly, we construct realizations of a stochastic vector variable $\mathbf{f}_k$ with factor scores. For this the matrix $\mathbf{\Phi}^{\frac{1}{2}}_{tk}$ is multiplied by a vector $\mathbf{x}_k$, which is a stochastic multivariate variable of order $q \times 1$, governed by $N(0,\mathbf{I})$. So we have $\mathbf{f}_k = \mathbf{\Phi}^{\frac{1}{2}}_{tk}\mathbf{x}_k$ and $\mathcal{E}(\mathbf{f}_k\mathbf{f}'_k) = \mathcal{E}(\mathbf{\Phi}^{\frac{1}{2}}_{tk}\mathbf{x}_k\mathbf{x}'_k\mathbf{\Phi}^{\frac{1}{2}}_{tk}) = \mathbf{\Phi}^{\frac{1}{2}}_{tk}\mathcal{E}(\mathbf{x}_k\mathbf{x}'_k)\mathbf{\Phi}^{\frac{1}{2}}_{tk} = \mathbf{\Phi}_{tk}$. The n_k

realizations of the stochastic multivariate variable $\mathbf{f}_k$ are put in a matrix $\mathbf{F}_k$ with factor scores. The simulated matrix $\mathbf{F}_k$ is not of interest in itself, but is indispensable in simulating other matrices, which are of interest in the present study.

The raw scores for sample k are now calculated as

$$\mathbf{Y}_k = \mathbf{F}_k\mathbf{P}'_{tk} + \mathbf{E}_k, \tag{3.2}$$

where $\mathbf{E}_k$ contains n_k realizations of the stochastic multivariate variable $\mathbf{e}_k$. There are now two distinct ways to proceed with the data simulation. When a covariance matrix is desired, it is computed directly from the matrix $\mathbf{Y}_k$ (the scores in $\mathbf{Y}_k$ are first centered to get deviation scores). When a correlation matrix is desired, the raw scores are standardized to obtain a matrix $\mathbf{Z}_k$.

Note that the present way of data simulation is exactly the opposite from the route taken by data analysis (i.e. factor analysis). In factor analysis the variable scores are – as well as possible – compressed into factor scores, while in the present simulation process the simulated factor scores are decompressed to arrive at variable scores.

For each generated matrix with scores on the variables, the true values are demanded for the weights matrix, the pattern matrix, the structure matrix and the matrix with correlations between the factors, denoted by $\mathbf{W}_{tk}$, $\mathbf{P}_{tk}$, $\mathbf{S}_{tk}$ and $\mathbf{\Phi}_{tk}$, respectively. Of these four matrices, the matrices $\mathbf{P}_{tk}$ and $\mathbf{\Phi}_{tk}$ are chosen before the simulation. The four true matrices are required because in the present study, these true values are to be compared with the values found with the several methods of analysis. It will now be described how these matrices are chosen or derived when a covariance matrix and a correlation matrix are simulated, respectively.

3.2.1 Choosing a true pattern matrix $\mathbf{P}_{tk}$ and defining a true structure matrix $\mathbf{S}_{tk}$ when simulating a correlation or a covariance matrix

The true pattern matrix is chosen to be simple. Therefore, the pattern matrix is chosen such that each variable is constructed from only one or two factors.

When simulating a population correlation matrix, the true pattern matrix $\mathbf{P}_{tk}$ and the error coefficients in the diagonal matrix $\mathbf{D}^2_{ek}$ may not be chosen freely, because the variances of the simulated variable scores are required to be one. That is, when

$$\mathbf{z}_k = \mathbf{P}_{tk}\mathbf{f}_k + \mathbf{e}_k, \tag{3.3}$$

where $\mathbf{e}_k$ is an m–dimensional random variable with a multivariate normal distribution $N(0,\mathbf{D}^2_{ek})$, we have

$$\mathcal{E}(\mathbf{z}_k\mathbf{z}_k') = \mathcal{E}(\mathbf{P}_{tk}\mathbf{f}_k\mathbf{f}_k'\mathbf{P}_{tk}') + \mathcal{E}(\mathbf{P}_{tk}\mathbf{f}_k\mathbf{e}_k') + \mathcal{E}(\mathbf{e}_k\mathbf{f}_k'\mathbf{P}_{tk}') + \mathcal{E}(\mathbf{e}_k\mathbf{e}_k')$$

$$= \mathbf{P}_{tk}\mathcal{E}(\mathbf{f}_k\mathbf{f}_k')\mathbf{P}_{tk}' + \mathcal{E}(\mathbf{e}_k\mathbf{e}_k') = \mathbf{P}_{tk}\mathbf{\Phi}_{tk}\mathbf{P}_{tk}' + \mathbf{D}^2_{ek} = \mathbf{R}_{tk}, \tag{3.4}$$

the population correlation matrix, and hence

$$\mathbf{D}^2_{ek} = \mathbf{I}\text{–diag}(\mathbf{P}_{tk}\mathbf{\Phi}_{tk}\mathbf{P}_{tk}'). \tag{3.5}$$

In order to simulate data for which (3.5) holds, we might choose the true pattern matrices at wish (provided that the diagonal elements of $(\mathbf{P}_{tk}\mathbf{\Phi}_{tk}\mathbf{P}_{tk}')$ do not exceed 1), and take the error distribution according to (3.5).

To get insight in the signal to noise ratio for the constructed patterns, it should be noted that the variance of the 'true' part of the variables is given on the diagonal of $\mathbf{P}_{tk}\mathbf{\Phi}_{tk}\mathbf{P}_{tk}'$ and that of the error part by the diagonal elements of matrix $\mathbf{D}_{ek}$. The ratios of corresponding elements give the signal to noise ratio.

The true structure matrix is calculated as $\mathbf{S}_{tk} = \mathbf{P}_{tk}\mathbf{\Phi}_{tk}$. When the matrix $\mathbf{P}_{tk}$ pertains to nonoverlapping clusters of variables, the nonzero values of the true pattern matrix return in matrix $\mathbf{S}_{tk}$. When the off–diagonal elements of the matrix $\mathbf{\Phi}_{tk}$ are nonzero, the matrix $\mathbf{S}_{tk}$ contains nonzero values in the places where the true pattern matrix contains zero values. As a consequence, in this situation the true pattern matrix is simple and the true structure matrix is not.

The present approach to data simulation is the same as the one used in many studies comparing Principal Components Analysis with Factor Analysis (e.g. Velicer & Jackson (1990), Velicer et al. (1982), Snook & Gorsuch (1989) and Widaman (1993)). In those studies, the true correlation matrix as it holds in the population is given by

$$\mathbf{R}_{tk} = \mathbf{P}_{tk}\Phi_{tk}\mathbf{P}'_{tk} + \mathbf{U}^2_{tk}, \tag{3.6}$$

where $\mathbf{U}^2_{tk}$ is a diagonal matrix containing unique variances of the variables in the population, ensuring that the matrix $\mathbf{R}_{tk}$ has ones on the diagonal. Definition (3.6) is, in fact, equal to (3.4). The true pattern matrix always contains nonoverlapping clusters of variables (called congeneric factors by Widaman, 1993) with loadings which are less than one (mostly the values .4, .6 and .8 are chosen). The difference in data construction of the studies mentioned with the present study is that in the present study scores are constructed for all (simulated) individuals in a sample, while in the studies mentioned above, sample covariance matrices are constructed. In the present study, besides similar situations, also some more complex situations are simulated.

3.2.2 The non-uniqueness of the true weights matrix $\mathbf{W}_{tk}$

It will now be shown that, contrary to the true pattern and the true structure matrix, it is not possible to define a true weights matrix other than by arbitrary choice.

The "true scores matrix" $\mathbf{F}_k\mathbf{P}'_{tk}$ can be defined according to (3.2) with zero error. The simulated factor scores matrix $\mathbf{F}_k$ should contain linear combinations of these true scores, where the weights for these linear combinations define "true weights". Now the complete class of true weights matrices $\mathbf{W}_{tk}$, which satisfy

$$\mathbf{F}_k\mathbf{P}'_{tk}\mathbf{W}_{tk} = \mathbf{F}_k \tag{3.7}$$

will be sought. Any weights matrix satisfying (3.7) can serve as a true weights matrix. After premultiplication of (3.7) with $(\mathbf{F}'_k\mathbf{F}_k)^{-1}\mathbf{F}'_k$ (where we assume rank($\mathbf{F}_k$)=q) we have

$$\mathbf{P}'_{tk}\mathbf{W}_{tk} = \mathbf{I}. \tag{3.8}$$

A general formula for the true weight matrices $\mathbf{W}_{tk}$ can be written as

$$\mathbf{W}_{tk} = \mathbf{P}_{tk}\mathbf{U}_k + \mathbf{P}^{\perp}_{tk}\mathbf{V}_k, \tag{3.9}$$

where the columns of $\mathbf{P}^{\perp}_{tk}$ are orthogonal to the columns of $\mathbf{P}_{tk}$, and $\mathbf{U}_k$ and $\mathbf{V}_k$ are arbitrarily chosen. It is important to note that the matrix $\mathbf{P}^{\perp}_{tk}$ is

of order $m\times(m-q)$ and the matrix $\mathbf{V}_k$ is of order $(m-q)\times q$. Substituting (3.9) into (3.8) gives

$$\mathbf{P}'_{tk}(\mathbf{P}_{tk}\mathbf{U}_k + \mathbf{P}^{\perp}_{tk}\mathbf{V}_k) = \mathbf{P}'_{tk}\mathbf{P}_{tk}\mathbf{U}_k = \mathbf{I}. \tag{3.10}$$

From this it follows that

$$\mathbf{U}_k = (\mathbf{P}'_{tk}\mathbf{P}_{tk})^{-1} \tag{3.11}$$

and the complete class of true weight matrices $\mathbf{W}_{tk}$ is now given by

$$\mathbf{W}_{tk} = \mathbf{P}_{tk}(\mathbf{P}'_{tk}\mathbf{P}_{tk})^{-1} + \mathbf{P}^{\perp}_{tk}\mathbf{V}_k. \tag{3.12}$$

With the derivation of (3.12) we have found that with a true scores matrix $\mathbf{F}_k\mathbf{P}'_{tk}$ of low rank q, any true weights matrix of full column rank q satisfying (3.12) will do. So there simply is no weights matrix which can be said to be more true than others calculated with (3.12). Because of this, no true weights matrix is present in the true model.

3.3 DEFINITION OF DATA CATEGORIES

Up till now, the simulation of scores of n_k individuals in sample k has been considered. Each sample is taken from a certain population, defined by a true pattern matrix $\mathbf{P}_{tk}$, a true structure matrix $\mathbf{S}_{tk}$ and a true correlation matrix $\mathbf{\Phi}_{tk}$. In the present study, the constructed data sets with (two or more samples) contain either samples taken from one population or samples from two or more (different) populations. These categories will be discussed in Sections 3.3.1 and 3.3.2, respectively. For both categories the percentages of explained variance, to be expected when simulated populations are analyzed with the three SCA-methods of analysis, will be discussed in Section 3.3.3. In the simulated populations the measurement error and sampling variability are excluded from the data. It will be shown that when sampling from different populations, the method SCA-W can, even when errorless data are analyzed, greatly outperform the methods SCA-P and SCA-S.

3.3.1 Sampling from one population

When samples are taken from one and the same population, the same true pattern matrix $\mathbf{P}_t$ and the same true correlation matrix $\mathbf{\Phi}_t$ underlies each sample to be simulated. So for each sample k we have $\mathbf{P}_{tk} = \mathbf{P}_t$ and $\mathbf{\Phi}_{tk} = \mathbf{\Phi}_t$. For this sort of data (from now on referred to as one–population data) all the samples have the same underlying true pattern matrix $\mathbf{P}_t$ and the same underlying true structure matrix $\mathbf{S}_t$. Furthermore, the true pattern matrix $\mathbf{P}_t$ underlying one–population data is simple (usually pertaining to nonoverlapping clusters of variables) while the true structure matrix $\mathbf{S}_t$ is usually not simple (unless the true matrix $\mathbf{\Phi}_t$ is the identity matrix).

A special case of one–population data is the situation where the true pattern matrices are not identical, but columnwise perfectly congruent across groups. In this situation, the amount of measurement error differs over the groups, but they can still be considered as coming from the same population, because it is obvious that the factors are still interpreted the same in all groups, although differing in strengths across groups.

To ensure that the columns of the two true structure matrices are perfectly congruent, it suffices to choose the matrix $\mathbf{\Phi}_{tk}$ as the identity matrix for all groups and the matrix $\mathbf{P}_{tk}$ columnwise proportional over the groups (because then we have $\mathbf{S}_t = \mathbf{P}_t$) or to choose the matrix $\mathbf{\Phi}_{tk}$ unequal to the identity matrix, but the same for all groups and choose the matrix $\mathbf{P}_{tk}$ either exactly the same for all groups (i.e. the error for each variable may differ, as long as all groups are treated equally) or choose the matrix $\mathbf{P}_{tk}$ so that within each simulated group all nonzero loadings are equal.

3.3.2 Sampling from two or more populations

When samples are taken from different populations, a different true pattern matrix $\mathbf{P}_{tk}$ is chosen for each sample to be simulated. The correlation matrix $\mathbf{\Phi}_{tk}$ may be chosen different for each sample to be

simulated, but this is not a necessity. For this reason, we will call these data two-or-more-populations data. The columns of the true pattern matrices are not proportional over groups, because then we would have one-population data. For two-or-more-populations data all the samples have different underlying true pattern matrices $\mathbf{P}_{tk}$ and different true structure matrices $\mathbf{S}_{tk}$. In the present study, the true pattern matrices $\mathbf{P}_{tk}$ underlying two-or-more-populations data are simple (nonoverlapping clusters of variables) while the true structure matrices $\mathbf{S}_{tk}$ are not simple (unless the matrix $\mathbf{\Phi}_{tk}$ is the identity matrix).

3.3.3 Explained variances in the populations

In Section 2.6.1, it was shown that the three methods of simultaneous components analysis SCA-W, SCA-P and SCA-S show a hierarchical order in the amounts of variance explained by the components found with the respective methods, which is written as

$$f_w \geq f_p \geq f_s, \tag{3.13}$$

where f stands for the variance explained and the indexes w, p and s stand for SCA-W, SCA-P and SCA-S, respectively. For the two categories of data presented, a stronger relationship between the amounts of variance explained by the three methods can be given when no error (or almost no error) has been added to the simulated scores. This will now be explained.

When one population is simulated with q factors, it is obvious that there is only one correlation matrix describing that population without error. This true correlation matrix for that population can be described as $\mathbf{R}_t = \mathbf{P}_t\mathbf{\Phi}_t\mathbf{P}_t' = \mathbf{S}_t\mathbf{\Phi}_t^{-1}\mathbf{S}_t'$. Therefore, SCA-P and SCA-S will both be able to explain exactly 100% of the variance when q components are drawn, because there is one pattern and one structure matrix perfectly describing both samples. From (3.13) it follows that SCA-W will also explain 100% of the variance. So for one-population data without error we have

$$f_w = f_p = f_s. \tag{3.14}$$

When two or more different populations are simulated the correlation

matrices describing the populations will differ to some degree. When the columns of the matrices $\mathbf{P}_{tk}$ span different spaces we have

$$f_w > f_p. \tag{3.15}$$

The proof for this strict inequality will now be given in two steps. Firstly, it will be shown that, in the error-free case, SCA-W is always able to explain 100% of the variance, whether we have one-population data or two-or-more-populations data. Secondly, it will be shown that SCA-P can not explain 100% of the variance in this case.

When population scores are simulated as described, the (errorless) data matrix $\mathbf{Z}_{tk}$ has at most rank q. Now, assume that a weights matrix $\mathbf{W}$ is used to calculate a factor scores matrix $\mathbf{F}_k = \mathbf{Z}_{tk}\mathbf{W}$, which has the same rank as the matrix $\mathbf{Z}_{tk}$. The columns of this factor scores matrix $\mathbf{F}_k$ span the same space as the columns of the scores matrix $\mathbf{Z}_{tk}$. Hence, with suitable pattern matrices $\mathbf{P}_k$, $k=1,...,p$, all the matrices $\mathbf{Z}_{tk}$ can be *perfectly* retrieved, so we have $\mathbf{Z}_{tk}\mathbf{W}\mathbf{P}_k' = \mathbf{Z}_{tk}$. This explains why always (whether we have one-population data or two-or-more-populations data) a weights matrix can be found, that is the same for all populations and has perfect fit in each population. So SCA-W will always be able to explain 100% of the variance.

It is, however, not possible to describe different true correlation matrices $\mathbf{R}_{tk}$ in two-or-more-populations data, which have at most rank q, with one pattern matrix, because the columns of the true pattern matrices $\mathbf{P}_{tk}$, making up the true correlation matrices $\mathbf{R}_{tk} = \mathbf{P}_{tk}\mathbf{\Phi}_{tk}\mathbf{P}_{tk}'$, span different subspaces. So the components found with SCA-P will always explain less than 100% of the variance.

The largest possible difference in explained variance between SCA-W and SCA-P, in the case where just two rank q correlation matrices are analyzed, is 50%. This happens when the true pattern matrices $\mathbf{P}_{t1}$ and $\mathbf{P}_{t2}$ are chosen orthogonal. Then the components found with SCA-W explain 100% of the variance and the components found with SCA-P explain 50% of the variance (see Appendix C).

For SCA-S a problem arises in this situation. When the matrices $\mathbf{R}_{tk}$, of low rank q, span different q dimensional column subspaces, it is impossible to define a q dimensional structure matrix $\mathbf{S}$ that is the same

for all $\mathbf{R}_{tk}$, because this matrix $\mathbf{S}$ has to satisfy $\mathbf{S} = \mathbf{R}_{tk}\mathbf{W}_k$. Therefore, the solution of SCA–S for these matrices $\mathbf{R}_{tk}$ is not defined and nothing can be said about the amount of variance explained by the solutions found with SCA–S. These solutions do not exist. This is not surprising because for SCA–S it was assumed that the matrix $\mathbf{R}_k$ is nonsingular, k=1,...,p.

3.4 A MEASURE FOR THE SIMILARITY OF POPULATIONS

It has been described above how samples of simulated data can be drawn from the same or from different populations. The amount of difference between populations can be manipulated by means of the choice of the true patterns. A question that arises is: "How can we quantify the difference between the populations that underlie the simulated samples?" To answer this question, a measure, called the "Congruence Measure" (CM), which quantifies the similarity of populations, was chosen.

3.4.1 The Congruence Measure

To assess to what degree the columns of a matrix $\mathbf{A}$ lead to the same interpretation as the columns of a matrix $\mathbf{B}$, Tucker's (1951) measure of congruence was adopted. The congruence between two vectors $\mathbf{a}$ and $\mathbf{b}$ (columns of the matrices $\mathbf{A}$ and $\mathbf{B}$) is defined as

$$\phi(\mathbf{a},\mathbf{b}) = (\mathbf{a}'\mathbf{a})^{-\frac{1}{2}}\mathbf{a}'\mathbf{b}(\mathbf{b}'\mathbf{b})^{-\frac{1}{2}}. \qquad (3.16)$$

When the congruence between two vectors of factor loadings on the same variables is higher than .85, the two factors will generally be judged to be more equal than unequal (see, for instance, Haven & Ten Berge, 1977). As a generalization of this congruence measure for two vectors to a congruence measure for two matrices with q columns each, first the congruence is calculated of each column $\mathbf{a}_l$, l = 1,...,q, of the matrix $\mathbf{A}$, with each column $\mathbf{b}_l$ of the matrix $\mathbf{B}$. This results in a $q \times q$ matrix $\mathbf{M}_\phi$ with q^2 congruences. As a second step, the highest mean of the absolute values of q congruences is taken, taking only one congruence from each row and each column of $\mathbf{M}_\phi$. This mean is called the Congruence Measure (CM). The

CM will be calculated for the true pattern or structure matrices of populations as a measure to quantify the similarity of two populations.