FACTOR ANALYSIS OF DICHOTOMIZED VARIABLES*

BY

ANDERS CHRISTOFFERSSON

UNIVERSITY OF UPPSALA

UPPSALA, SWEDEN

An approach for multiple factor analysis of dichotomized variables is presented. It is based on the distribution of the first and second order joint probabilities of the binary scored items. The estimator is based on the generalized least squares principle. Standard errors and a test of the fit of the model is given.

Introduction

Most methods for factor analysis are based on the sample-correlation matrix and thus implicitly assume that the variables under study are measured on at least an interval scale. A common practice for the case of dichotomized variables is to apply some method for factor analysis on the tetrachoric correlations. This has sometimes resulted in a break down because the methods usually require (among other things) that the correlation matrix is Grammian.

For the case of one factor, there are two different maximum likelihood approaches, both taking into account that the variables are dichotomized: i) The conditional maximum likelihood method [Lord 1968].


The first method estimates the threshold levels (cutting point), the factor loadings and the factor scores simultaneously. The validity of this approach "has not been rigorously proven . . ." [Lord 1970]. Intuitively, this seems to be impossible because the approach is equivalent to estimating factor loadings, factor scores and residual variances simultaneously in the case of interval measurements. This is not possible with the maximum likelihood method [Anderson and Rubin 1956] unless some further conditions are imposed, such as assuming that the residual variances are pairwise the same. Actually it is possible to construct models for which the maximum of the conditional likelihood does not correspond to the true parameter set (not even the loadings). The unconditional maximum likelihood approach, on the other hand, is theoretically optimal (under certain conditions). How-

* This work is supported by The Bank of Sweden Tercentenary Fund
ever, in many practical applications, it cannot be used because "... computational difficulties limit the solution to not more than 10 or 12 items in any one analysis—a number too small for typical psychological test applications". [Bock and Lieberman 1970].

The approach which is presented in this paper is based on the marginal distributions of single and pairs of items. This involves of course a loss of information compared with the unconditional maximum likelihood method. However, this loss seems to be of little practical importance. The merits of this approach are that it can handle larger amounts of data and that it is possible to deal with more complicated models, i.e. unrestricted and restricted multifactor models and the model for analysis of covariance structures given by Jöreskog [1970].

1. Factor Model

Assume that the latent M-dimensional variable $y$ allows the usual $K$-factor model

$$y = \mu + \Lambda \phi + \delta$$

where $\Lambda$ is the $M \times K$ matrix of factor loadings, $\phi$ is the $K$-dimensional vector of factors and $\delta$ is the $M$-dimensional residual vector, distributed independently of $\phi$. $\mu$ denotes the expectation of $y$ i.e. $E(y) = \mu$. We further assume that the expectation of $\phi$ and $\delta$ is zero and that $y$ has a multivariate normal distribution. It then follows that the covariance matrix of $y$ is

$$E[(y - \mu)(y - \mu)'] = \Sigma = \Lambda \Phi \Lambda' + \psi$$

where $\Phi$ is the covariance matrix of the factors and $\psi$ is the covariance matrix of $\delta$. $\psi$ is assumed to be diagonal, with nonnegative elements.

To make this decomposition unique, i.e. to make the model identified, we must impose at least $K^2$ independent conditions. In the unrestricted case these conditions are usually imposed in a way suitable for the estimation method that is used. The estimates are then transformed (rotated) according to some criteria, for example the varimax criteria. In the restricted case, the restrictions are given by the nature of the application. In this paper we will not treat the problem of identifiability. We will only be concerned with the problem of estimation of parameters in identified models.

Suppose now that the manifest variable $y^*$ is such that its $i$-th component $y_{i}^*$ equals 1 if the corresponding component of $y$, $y_i$, is greater than or equal to a number $h_i$, called threshold level, and otherwise zero. That is

$$y_{i}^* = \begin{cases} 1 & \text{if } y_i \geq h_i \\ 0 & \text{if } y_i < h_i \end{cases}$$

The distribution of $y^*$ is of the type