FACTOR ANALYSIS AS MATRIX DECOMPOSITION

JAN DE LEEUW

ABSTRACT. Meet the abstract. This is the abstract.

1. INTRODUCTION

Suppose we have n measurements on each of taking m variables. Collect these measurements in an $n \times m$ matrix Y. In this introductory section we briefly review two classical factor analysis models, more precisely linear common factor models. For a more comprehensive discussion we refer to Anderson and Rubin [1956] and to Anderson [1984].

In common factor analysis we suppose that Y is the sum of a *common part* and a *unique part*. This is analogous to discussing data as composed of a signal and a noise part (or a fit and error part) in other data analysis contexts. We write the model, informally, in algebraic form[1] as

$$\begin{aligned}
\underset{n\times m}{Y} &= \underset{n\times m}{F} + \underset{n\times m}{U},\\
\underset{n\times m}{F} &= \underset{n\times p}{H}\,\underset{p\times m}{A},\\
\underset{n\times m}{U} &= \underset{n\times q}{E}\,\underset{q\times m}{D}.
\end{aligned}$$

Thus the common part consists of linear combinations of p common factors, and the unique part of linear combinations of q unique factors. The model basically says that both common and unique parts are of low rank. There is also a notion that the common and unique parts are orthogonal in some sense.

Date: April 6, 2008 — 19h 23min — Typeset in TIMES ROMAN.

2000 *Mathematics Subject Classification.* 00A00.

Key words and phrases. Binomials, Normals, LaTeX.

[1] Observe that we show the dimensions of a matrix by giving the numbers of rows and columns under the symbol of the matrix.

2. Factor Analysis Models

There are various ways in which the general idea of factor analysis can be made more precise by formulating it as an explicit statistical model.

2.1. **Random Factor Model.**

The matrix Y is supposed to be a realization[2] of a matrix-valued random variable $\underline{Y}$.

In random score factor analysis we assume that the random variable $\underline{Y}$ has a random common part $\underline{F}$ and a random unique part $\underline{U}$. Thus

$$\underset{n\times m}{\underline{Y}} = \underset{n\times m}{\underline{F}} + \underset{n\times m}{\underline{U}} .$$

The common part is a linear combination of a number, say p, of common factors $\underline{H}$, i.e.

$$\underset{n\times m}{\underline{F}} = \underset{n\times p}{\underline{H}} \underset{p\times m}{A'} .$$

The unique part is a linear combination of a number, say q, of unique factors $\underline{E}$.

$$\underset{n\times m}{\underline{U}} = \underset{n\times q}{\underline{E}} \underset{q\times m}{D'} .$$

The rows of $\underline{Y}$, corresponding with the different individuals, are assumed to be independent. Moreover we assume the specific parts are uncorrelated with the common factors, and with the other specific parts. For simplicity we assume all variables are centered, i..e have expectation zero.

2.2. **Fixed Factor Model.**

This, at least, is what the *random factor model* says. There is also a *fixed factor model*, which assumes

$$\underline{Y} = F + \underline{E}.$$

Now the common part is a bilinear combination of a number of common factor loadings a_{js} and common factor scores u_{is}, i.e.

$$F = UA'.$$

In the fixed model we merely assume the specific parts are uncorrelated with the other specific parts.

[2] We use the convention of underlining random variables.

2.3. **Mixed models.** The random factor model explained above was criticized soon after it was formally introduced by Lawley.

> The point is that in factor analysis different individuals are regarded as drawing their scores from *different* k-way distributions, and in these distributions the mean for each test is the true score of the individual on that test. Nothing is implies about the distribution of observed scores over a population of individuals, and one makes assumptions only about the error distributions [Young, 1940, pag. 52].

This quotation suggests the random factor model $\underline{Y} = \underline{U}A' + \underline{E}$, where $\mathbf{E}(\underline{U}) = U$. Let $\underline{H} = \underline{U} - U$, then $\underline{Y} = UA' + \underline{H}A' + \underline{E}$.

2.4. **Bilinear Models.**

2.5. **Covariance Form.** and it follows that if $\Sigma = \mathbf{E}(\underline{yy}')$ then

$$\Sigma = A\Omega A' + \Delta^2,$$

where $\Omega = \mathbf{E}(\underline{tt}')$, and where Δ^2, the covariance matrix of the specifics, is diagonal. We see that we can formulate the model as one in which a number random variables are decomposed, and as a parametric model for the covariance matrix of these random variables.

The next step is to connect the population model to observations following the model. The standard approach is to assume we have n repeated independent trials.

2.6. **Data Model.** So far, we have looked at models of the form $\underline{Y} = \underline{F} + \underline{E}$ or $\underline{Y} = F + \underline{E}$. This follows the standard statistical practice of embedding our data in a replication framework by assuming they are realizations of random variables. We then make calculations on and statements about the random variables, hoping that we will get back to our data eventually.

Instead of taking this circuitous route, we can also look directly at $Y = F + E$, with $F = UA'$. and $U'U = I, E'E = D^2$, and $U'E = 0$. This formulates the model directly in terms of the data. We cannot falsify the model by standard statistical methods, because there is no replication framework, but the general notions of stability and fit continue to apply.

3. Estimation

3.1. **Covariance Matrix Methods.** The dominant estimation method in factor analysis is multinormal maximum likelihood for the random factor model. It was first proposed by Lawley [1939], and then popularized and programmed by Jöreskog [1967]. The negative log-likelihood measure the distance between the sample and population covariance model, and we must minimize

$$\mathscr{L}(A,D) = n\log|\Sigma| + n\ \mathbf{tr}\ \Sigma^{-1}S,$$

with S the sample covariance matrix of Y, and with $\Sigma = A\Omega A' + \Delta^2$.

In Anderson and Rubin [1956] the impressive machinery developed by the Cowles Commission was applied to both the fixed and random factor analysis model. Maximum likelihood was applied to the likelihood function of the covariance matrix, assuming multivariate normality.

3.2. **Data Matrix Methods.** Lawley's maximum likelihood procedures were criticized soon after they appeared by Young [1940].

> Such a distribution is specified by the means and variances of each test and the covariances of the tests in pairs; it has no parameters distinguishing different individuals. Such a formulation is therefore inappropriate for factor analysis, where factor loadings of the tests and of the individuals enter in a symmetric fashion in a bilinear form [Young, 1940, pag. 52].

Young proposed to minimize the log-likelihood of the data

$$\mathscr{L} = n\log|D| + m\log|E| + \mathbf{tr}\ (Y-UA')'E^{-1}(Y-UA')D^{-1}$$

where D and E are *known* diagonal matrices with column (variable) and row (individual) weights. The solution is given by a weighted singular value decomposition of Y.

The basic problem with Young's method is that it supposes the weights to be known. One solution, suggested by Lawley [1942], is to estimate them along with the loadings and scores. In the more common case, in which there are no person-weights, we have

$$\mathscr{L}(U,A,D) = n\log|D| + \mathbf{tr}\ (Y-UA')D^{-1}(Y-UA')'.$$

Lawley suggests to alternate minimization over (U,A), which is done by weighted singular value decomposition, and minimization over diagonal D, which simply amounts to computing the average sum of squares of the residuals for each variable. Iterating these two minimizations produces an algorithm intended to minimize the likelihood. But unfortunately it does not work.

> A rather disconcerting feature of the new method is, however, that iterative numerical solutions of the estimation equations either fail to converge, or else converge to unacceptable solutions in which one of more of the measurements have zero error variance. It is apparently impossible to estimate scale as well as location parameters when so many unknowns are involved [Whittle, 1952, pag. 224]

In fact, if we look at the loss function we can see it is unbounded below. We can choose loadings to fit one variable perfectly, and then let the corresponding variance term approach zero [Anderson and Rubin, 1956].

Several other remedies have been proposed to rescue the weighted least squares methods. Whittle [1952] suggested to take D proportional to the variances of the variables. This amounts to doing a singular value decomposition of the standardized variables. Jöreskog [1962] makes the more reasonable choice of setting D proportional to the reciprocals of the diagonals of the inverse of the covariance matrix of the variables (i.e. to the residual variances when regressing each variable on the others). Of course in these approaches the weights themselves depend on the data Y, which means that simple weighted least squares theory does not apply.

An original approach was suggested by McDonald [1979]. Also see Etezadi-Amoli and McDonald [1983]. He proposes to maximize the determinant of the correlation matrix of the matrix of residuals $R = Y - UA'$. This criterion can be derived by using the fact that if we minimize over diagonal D, then

$$\min_D \mathscr{L}(U,A,D) = n\log|\mathbf{diag}R'R|,$$

while if we minimize over unrestricted S we have

$$\min_S \mathscr{L}(U,A,S) = n\log|R'R|,$$

The difference of the two is the logarithm of the determinant of the correlation matrix of the residuals. The approach is clearly scale-free, and the maximum of zero is attained if we can make the residuals exactly uncorrelated. Computational and statistical properties of this so-called *maximum likelihood ratio method* are quite complicated, however.

4. A NEW APPROACH

In this paper we introduce a new approach to simultaneous least squares estimation of both loadings and scores, which moves away even further from the standard Lawley-Joreskog approach, and is formulated even more directly in terms of the data.

We simply write

$$Y = UA' + ED,$$

and we minimize

$$\sigma = \mathbf{tr}(Y - UA' - ED)'(Y - UA' - ED)$$

over (U,E,A,D), requiring $U'U = I, E'E = I$, and $U'E = 0$.

The method may seem to be quite similar to the approach proposed by Paul Horst in his book [Horst, 1965]. Where we differ from Horst is in our additional assumptions that D is diagonal and that E has the same size as the data Y. This puts us solidly in the common factor analysis framework. Horst, on the contrary, only makes the assumption that there is a small number of common and residual factors, and he then finds them by truncating the singular value decomposition. Separating common and unique factors can be done later by using rotation techniques. For Horst factor analysis is just principal component analysis with some additional interpretational tools.

The algorithm to minimize our loss function is of the alternating least squares type. We start with an initial estimate $A^{(0)}$ and $D^{(0)}$ and then alternate

$$\begin{aligned}
\left[U^{(k)} \quad | \quad E^{(k)}\right] &\in \mathbf{procrustus}\left[YA^{(k)} \quad | \quad YD^{(k)}\right],\\
A^{(k+1)} &= Y'U^{(k)},\\
D^{(k+1)} &= \mathbf{diag}(Y'E^{(k)}).
\end{aligned}$$

As shown in Appendix A, the Procrustus transformation is a set of matrices. We choose any one of its elements.

5. Block Factor Analysis

The analysis in the previous se3ction can be extended easily to a factor analysis of sets of variables. Suppose Y_j are $n \times m_j$ matrices, normalized such that $Y_j'Y_j = I$. The loss function we minimize is

$$\sigma = \sum_{j=1}^{m} \mathbf{tr}\ (Y_j - E_0A_j' - E_jD_j)'(Y_j - E_0A_j' - E_jD_j)$$

for all A_j and D_j, and over all E_j with $E_j'E_\ell = \delta^{j\ell}I$ for all j and ℓ.

If we define the matrices

$$Y = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_m \end{bmatrix},$$

$$E = \begin{bmatrix} E_0 & E_1 & \cdots & E_m \end{bmatrix},$$

$$A' = \begin{bmatrix} A_1' & A_2' & \cdots & A_m' \\ D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_m \end{bmatrix},$$

then the loss is

$$\sigma = \mathbf{tr}\ (Y - EA')'(Y - EA'),$$

where $E'E = I$.

APPENDIX A. AUGMENTED PROCRUSTUS

Suppose X is an $n \times m$ matrix of rank r. Consider the problem of maximizing $\mathbf{tr}\ U'X$ over the $n \times m$ matrices U satisfying $U'U = I$. This is known as the *Procrustus* problem, and it is usually studied for the case $n \geq m = r$. We want to generalize to $n \geq m \geq r$. For this, we use the singular value decomposition

$$X = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times r}{\Lambda} & \underset{r\times(m-r)}{0} \\ \underset{(n-r)\times r}{0} & \underset{(n-r)\times(m-r)}{0} \end{bmatrix} \begin{bmatrix} \underset{r\times m}{L_1'} \\ \underset{(m-r)\times m}{L_0'} \end{bmatrix}.$$

Theorem A.1. *The maximum of* $\mathbf{tr}\ U'X$ *over* $n \times m$ *matrices* U *satisfying* $U'U = I$ *is* $\mathbf{tr}\ \Lambda$, *and it is attained for any* U *of the form* $U = K_1L_1' + K_0VL_0'$, *where* V *is any* $(n-r) \times (m-r)$ *matrix satisfying* $V'V = I$.

Proof. Using a symmetric matrix of Lagrange multipliers leads to the stationary equations $X = UM$, which implies $X'X = M^2$ or $M = \pm(X'X)^{1/2}$. It also implies that at a solution of the stationary equations $\mathbf{tr}\ U'X = \pm\mathbf{tr}\ \Lambda$. The negative sign corresponds with the minimum, the positive sign with the maximum.

Now

$$M = \begin{bmatrix} \underset{m\times r}{L_1} & \underset{m\times(m-r)}{L_0} \end{bmatrix} \begin{bmatrix} \underset{r\times r}{\Lambda} & \underset{r\times(m-r)}{0} \\ \underset{(m-r)\times r}{0} & \underset{(m-r)\times(m-r)}{0} \end{bmatrix} \begin{bmatrix} \underset{r\times m}{L_1'} \\ \underset{(m-r)\times m}{L_0'} \end{bmatrix}.$$

If we write U in the form

$$U = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times m}{U_1} \\ \underset{(n-r)\times m}{U_0} \end{bmatrix}$$

then $X = UM$ can be simplified to

$$U_1L_1 = I,$$
$$U_0L_1 = 0,$$

with in addition, of course, $U_1'U_1 + U_0'U_0 = I$. It follows that $U_1 = L_1'$ and

$$\underset{(n-r)\times m}{U_0} = \underset{(n-r)\times(m-r)}{V} \underset{(m-r)\times m}{L_0'},$$

with $V'V = I$. Thus $U = K_1L_1' + K_0VL_0'$. □

Appendix B. The Fundamental Theorem of Factor Analysis

There is a closely related theorem which is known, or used to be known, as the fundamental theorem of factor analysis. It took the cumulative efforts of many fine minds, starting with Spearman, about 25 years to come up with a proof of this theorem. The fact that it follows easily from the singular value decomposition shows the power of modern matrix algebra tools.

Theorem B.1. *Suppose* $\underset{n\times m}{X}$ *and* $\underset{m\times p}{A}$ *are such that* $X'X = AA'$*. Then there is an* $\underset{n\times p}{U}$ *such that* $U'U = I$ *and* $X = UA'$*.*

Proof. From $X'X = AA'$ we know that A has singular value decomposition

$$A = \begin{bmatrix} \underset{m\times r}{L_1} & \underset{m\times(m-r)}{L_0} \end{bmatrix} \begin{bmatrix} \underset{r\times r}{\Lambda} & \underset{r\times(p-r)}{0} \\ \underset{(m-r)\times r}{0} & \underset{(m-r)\times(p-r)}{0} \end{bmatrix} \begin{bmatrix} \underset{r\times p}{V_1'} \\ \underset{(p-r)\times p}{V_0'} \end{bmatrix},$$

where $r \leq p$ is the rank of both X and A. Observe that the left singular vectors of A are the right singular vectors of X.

Now we still have to solve $X = UA'$. Write

$$U = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times p}{U_1} \\ \underset{(n-r)\times p}{U_0} \end{bmatrix}.$$

Then $X = UA'$ simplifies to

$$I = U_1 V_1,$$
$$0 = U_0 V_1,$$

with in addition, of course, $U_1'U_1 + U_0'U_0 = I$. It follows that $U_1 = V_1'$ and

$$\underset{(n-r)\times p}{U_0} = \underset{(n-r)\times(p-r)}{W} \underset{(p-r)\times p}{V_0'},$$

with $W'W = I$. Thus $U = K_1 V_1' + K_0 W V_0'$. □

References

T.W. Anderson. Estimating linear statistical relationships. *The Annals of Statistics*, 12(1):1–45, 1984.

T.W. Anderson and H. Rubin. Statistical inference in factor analysis. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5*, pages 111–150. University of California Press, 1956.

J. Etezadi-Amoli and R.P. McDonald. A second generation nonlinear factor analysis. *Psychometrika*, 48:315–342, 1983.

P. Horst. *Factor Analysis of Data Matrices*. Holt, Rinehart and Winston, 1965.

K. G. Jöreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32:443–482, 1967.

K.G. Jöreskog. On the statistical treatment of residuals in factor analysis. *Psychometrika*, 27(4):335–354, 1962.

D.N. Lawley. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60:64–82, 1939.

D.N. Lawley. Further investigations in factor estimation. *Proceedings of the Royal Society of Edinburgh*, 61:176–185, 1942.

R.P. McDonald. The simultaneous estimation of factor loadings and scores. *British Journal of Mathematical and Statistical Psychology*, 32:212–228, 1979.

P. Whittle. On principal components and least squares methods in factor analysis. *Skandinavisk Aktuarietidskrift*, 35:223–239, 1952.

G. Young. Maximum likelihood estimation and factor analysis. *Psychometrika*, 6 (1):49–53, 1940.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu