Factor Analysis, Correspondence Analysis, ANOVA

Jan de Leeuw

First created on June 20, 2022. Last update on June 27, 2022

Abstract

The abstract

1 Introduction

1.1 Notation

We use the "Dutch Convention" of underlining random variables (Hemelrijk (1966)), to distinguish them from fixed quantities. This convention is particularly useful in the context of factor analysis, as we shall see in the rest of the paper. The symbol $\sim$ is used to indiacte the distribution of a random variable. Thus, for example, $\underline{z} \sim \mathcal{N}(\mu, \Sigma)$ says that the vector $\underline{x}$ has a normal distribution with mean μ and dispersion matrix Σ. We use the abbreviations "iff" for "if and only if", and "iid" to indicate that a number of random variables are "independent and identically distributed". An "rv" is a random variable. Capital letters are used for matrices, A' is the transpose of A, and A^+ its Moore-Penrose inverse. For parameters in statistical models we use Greek letters, keeping in mind that matrix parameters and thus Greek Capitals, which are the same as Roman capitals in many cases. We use := for definitions.

1.2 Terminology

To explain some terminology we use the algebraic version of common factor analysis (CFA). No rv's at all. The $n \times m$ data matrix X has measurements of n objects on m variables. In classical psychometrics the objects are individuals, the variables are tests. In CFA we look for a matrix decompositon $X = AB' + ED$, where A is $n \times p$, B is $m \times p$, E is $n \times m$, and D is $m \times m$. In addition we require that $E'E = I$, $A'A = I$, $E'A = 0$, and that D is diagonal. CFA is thus what is known these days as Exploratory Factor Analysis with Orthogonal Common Factors. We do not discuss Confirmatory Factor Analysis and Correlated Common Factors. It follows from the assumptions that $E'X = D$ and $A'X = B'$, and that $X'X = BB' + D^2$. It also follows that we must have $p \leq m \leq n$ for the decomposition to make sense.

Matrix A contains the *factor scores* or *common factor scores* of the objects, and B the *factor loadings* of the variables. E has the *unique factor scores* of the objects. The diagonal elements of D^2 are the *uniquenesses* and the diagonal elements of $BB' = XX - D^2$ are the *communalities* of the variables.

2 History

2.1 Lawley (1940)

In the 1930s leading statisticians such as Bartlett, Kendall, and Wishart, and excellent mathematicians such as Wilson, Piaggio, and Lederman, contributed to the analysis and development of common factor analysis. Lawley (1940), however, was the first to apply the awesome machinery of modern mathematical statistics, matrix algebra, and computation. He proposed a statistical model and corresponding estimation method for factor analysis that is, despite numerous generalizations and computational improvements, still basically the norm. So that is where we start our history.

++++++++ stochastization

$$\underline{x}_i = \mu + A\underline{\beta}_i + \underline{\epsilon}_i \tag{1}$$

The $\underline{z}_i$ and $\underline{e}_i$ are iid and jointly multinormal with means and dispersions

$$\begin{bmatrix} \underline{f}_i \\ \underline{e}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I & 0 \\ 0 & \Delta^2 \end{bmatrix} \right) \tag{2}$$

It follows that the $\underline{x}_i$ are iid and multinormal distribution as well, with

$$\underline{x}_i \sim \mathcal{N}(\mu, \Sigma), \tag{3}$$

with $\Sigma := AA' + \Delta^2$.

Maximum likelihood $\log\det\Sigma + \text{tr}\ \Sigma^{-1}C$

One remarkable property of equation .. is that there are no parameters for individuals. Or, to put it differently, are no factor scores. Describe the pop, thos asymmetry, factor score procedures

The joint distribution of $\underline{x}_i$ and $\underline{z}_i$ is

$$\begin{bmatrix} \underline{x}_i \\ \underline{z}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} AA' + \Delta^2 & A \\ A' & I \end{bmatrix} \right), \tag{4}$$

and consequently we have the conditional distributions

$$\underline{x}_i \mid \underline{z}_i = z_i \sim \mathcal{N}(\mu + A(AA' + \Delta^2)^{-1} z_i, \Delta^2) \tag{5}$$

$$\underline{z}_i \mid \underline{x}_i = x_i \sim \mathcal{N}(A\Sigma^{-1}(x - \mu),). \tag{6}$$

2.1.1 Young (1940)

Gale Young introduced psychometricians to low-rank matrix approximation (Eckart and Young (1936)) and to recovering coordinates of points in Euclidean space from matrices of distances (Young and Householder (1938)). In Young (1940) he criticized the basic factor analysis model used by Lawley (1940). We quote from his discussion section (Young (1940), page 52).

> A number of writers on the factor problem, e.g., Lawley, have formulated it as if different individuals taking a set of k tests were merely drawing samples from the same k-way distribution of variables in normal correlation. 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 symmetric fashion in a bi-linear form. It would perhaps be more suitable for psychophysics, where the differences between individuals are ignored and attention is focused on the test objects presented to them.
>
> 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 implied about the distribution of observed scores over a population of individuals, and one makes assumptions only about the error distributions.

Assume independent random samples of size one from n different probability distributions

$$\begin{aligned}\underline{x}_{ij} &= \sum_{s=1}^p z_{is}a_{js} + \underline{\epsilon}_{ij}\\ \underline{\epsilon}_{ij} &\sim \mathcal{N}(0,\sigma_{ij}^2)\end{aligned}$$

In addition all nm $\underline{\epsilon}_{ij}$ are independent, and all σ_{ij}^2 are supposed known. Note the symmetry between objects and variables in this model.

The negative log likelihood is

$$\mathcal{L}(Z,A) = \sum_{i=1}^n\sum_{j=1}^m \log\sigma_{ij}^2 + \sum_{i=1}^n\sum_{j=1}^m \frac{(x_{ij}-\sum_{s=1}^p z_{is}a_{js})^2}{\sigma_{ij}^2} \tag{7}$$

Because the σ_{ij}^2 are supposed known the first term is a constant, and maximum likelihood estimates can be found by minimizing the second term, a weighted sum of squares. In general, there is no closed form solution. We would now use alternating least squares or majorization to minimize ... without giving it too much thought. Young does analyze the special case in which $\sigma_{ij}^2 = \eta_i^2\xi_j^2$, with both η and ξ known. Then

$$\sum_{i=1}^n\sum_{j=1}^m \frac{(x_{ij}-\sum_{s=1}^p z_{is}a_{js})^2}{\sigma_{ij}^2} = \sum_{i=1}^n\sum_{j=1}^m(\eta_i^{-1}\xi_j^{-1}x_{ij} - \sum_{s=1}^p(\eta_i^{-1}z_{is})(\xi_j^{-1}a_{js}))^2, \tag{8}$$

and we can apply Eckart-Young to the scaled x_{ij} and then unscale the z_{is} and a_{js}. This obviously has the special cases $\sigma_{ij}^2 = \xi_j^2$ and $\sigma_{ij}^2 = \eta_i^2$ and $\sigma_{ij}^2 = \sigma^2$.

In fact, if $\sigma_{ij}^2 = \sigma^2$ it is not even necessary that σ^2 is known (Young (1940), page 50).

$$\mathcal{L}(Z, A, \sigma^2) = nm \log \sigma^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \sum_{s=1}^{p} z_{is} a_{js})^2 \tag{9}$$

We just compute the Eckart-Young approximation of X to get maximum likelihood estimates of „, and estimate

Young .. admits that in most situations

2.1.2 Lawley (1941)

Lawley reacted quickly and constructively to the criticism in Young (1940). In Lawley (1941) page 176-177

> In his solution of the factor problem Gale Young makes, for practical reasons, what is equivalent to the supposition that the error variance on a hypothetical infinity of trials is independent of the individual but depends only on the test. He assumes, however, that the error variances have been previously determined from other data. In the present paper we shall show that this last assumption is not strictly necessary since, given the scores of a group of individuals on a set of tests, it is in theory possible to estimate simultaneously not only the factor measurementsof the individuals and the factor loadings of the tests but also the error variances.

$$\underline{x}_{ij} = \mu_j + \sum_{s=1}^{p} a_{is} b_{js} + \underline{e}_{ij} \tag{10}$$

$$\underline{\epsilon}_{ij} \sim \mathcal{N}(0, \sigma_j^2)$$

$$\underline{x}_i \sim \mathcal{N}(\mu + B a_i, \Delta^2) \tag{11}$$

$$\mathcal{L}(Z, A) = n \sum_{j=1}^{m} \log \sigma_j^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(x_{ij} - \sum_{s=1}^{p} z_{is} a_{js})^2}{\sigma_j^2}$$

Lawley does not give his opinion about which model, ... or ..., is better. Unlike Young. I have no such opinion either, I just work here.

2.2 Burt (1947)

2.3 Whittle (1952)

2.4 Howe (1955)

2.5 Anderson and Rubin (1956)

2.6 Bargmann (1957)

2.7 Jöreskog (1962)

2.8 Gollob (1968)

2.9 De Leeuw (1973)

2.10 McDonald (1979)

3 HBM Estimation in FFA

Minus one half the log likelihood

$$\mathcal{L}_0(\theta,\Sigma) := n\log\det(\Sigma) + n\ \text{tr}\ \Sigma^{-1}C(\theta)$$
$$\mathcal{L}_0^\star(\theta) := \min_\Sigma \mathcal{L}_0(\theta,\Sigma) = \log\det(C(\theta)) + n$$

$$\hat{\Sigma} = C(\theta) := \frac{1}{n}R(\theta)'R(\theta)$$
$$\mathcal{L}_1(\theta,\Gamma) := n\log\det(\Gamma) + \sum_{i=1}^n r_i(\theta)'\Gamma^{-1}r_i(\theta)$$
$$\mathcal{L}_1^\star(\theta) := \min_\Gamma \mathcal{L}_1(\theta,\Gamma) = \log\det\ D(\theta) + n$$

$$\Delta^\star(\theta) := \mathcal{L}_1^\star(\theta) - \mathcal{L}_0^\star(\theta) = -\log\det\ \Gamma(\theta)$$
$$\Delta(\theta,\Gamma) = n\log\det(\Gamma) + n\text{tr}\ \Gamma^{-1}C(\theta) - n\log\det(C(\theta)) + n$$

= ## More general models

3.1 More general covariance structures

4 HBM Estimation in GYW

$$\mathcal{L}_0(\theta, \Sigma) = \sum_{i=1}^{n}\sum_{j=1}^{m} \log \sigma_{ij}^2 + \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{r_{ij}^2(\theta)}{\sigma_{ij}^2}$$

$$\mathcal{L}_0^\star(\theta) = \min_{\Sigma} \mathcal{L}_0(\theta, \Sigma) = \sum_{i=1}^{n}\sum_{j=1}^{m} \log r_{ij}^2(\theta) + nm$$

$$\mathcal{L}_1(\theta, \xi, \eta) = n\sum_{j=1}^{m} \log \eta_j^2 + m\sum_{i=1}^{n} \log \xi_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{r_{ij}^2(\theta)}{\xi_i^2 \eta_j^2}$$

$$\mathcal{L}_1^\star(\theta) = \min_{\eta,\xi} \ \mathcal{L}_1(\theta, \xi, \eta)$$

Now compute

$$\Delta^\star(\theta) := \mathcal{L}_1^\star(\theta) - \mathcal{L}_0^\star(\theta)$$

Define

$$\Delta(\theta, \xi, \eta) := \sum_{i=1}^{n}\sum_{j=1}^{m} \left\{ \frac{r_{ij}^2(\theta)}{\eta_i^2 \xi_j^2} - \log \frac{r_{ij}^2(\theta)}{\eta_i^2 \xi_j^2} - 1 \right\}$$

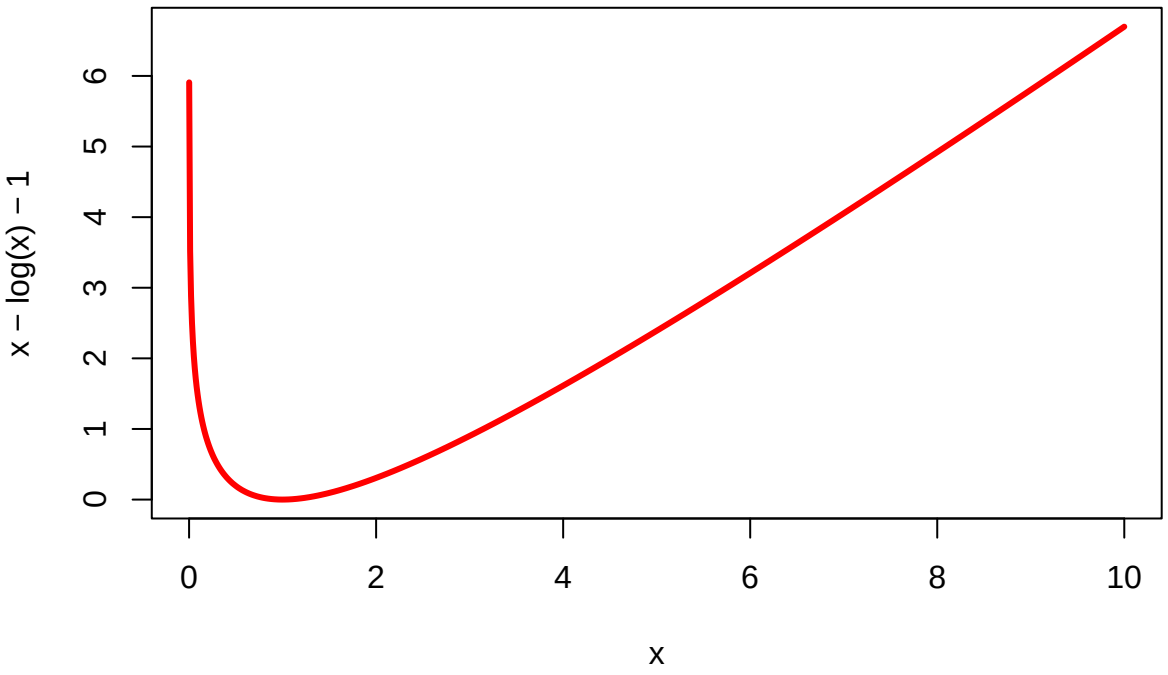
$$\min_{\theta} \Delta^\star(\theta) = \min_{\theta} \min_{\eta,\xi} \Delta(\theta, \xi, \eta)$$

Now $x - \log x - 1$ is a non-negative convex function which is zero if and only if $x = 1$. Thus proper loss function aiming for

$$\frac{r_{ij}^2(\theta)}{\eta_i^2 \xi_j^2} \approx 1$$

for all i and j.

```
x <- seq(.001,10, length = 1000)
plot(x, x - log(x) - 1, type = "l", col = "RED", lwd = 3)
```



$$r_{ij}(A,B) = x_{ij} - \sum_{s=1}^{p} a_{is} b_{js}$$

$$\frac{\partial r_{ij}(A,B)}{\partial a_{ks}} = -\delta^{ik} b_{js}$$

$$\frac{\partial \Delta}{\partial a_{is}} = -2 \sum_{j=1}^{m} \left\{ \frac{r_{ij}(A,B)}{\eta_i^2 \xi_j^2} - \frac{1}{r_{ij}(A,B)} \right\} b_{js}$$

$$\frac{\partial \Delta}{\partial b_{js}} = -2 \sum_{i=1}^{n} \left\{ \frac{r_{ij}(\theta)}{\eta_i^2 \xi_j^2} - \frac{1}{r_{ij}(\theta)} \right\} a_{is}$$

$$\frac{\partial^2 \Delta}{\partial a_{is} \partial a_{kt}} = 2\delta^{ik} \sum_{j=1}^{m} \left\{ \frac{1}{\eta_i^2 \xi_j^2} + \frac{1}{r_{ij}^2(A,B)} \right\} b_{js} b_{jt}$$

step 2

$$\frac{\partial \mathcal{L}}{\partial \eta_j^2} = n \frac{1}{\eta_j^2} - \frac{1}{\eta_j^4} \sum_{i=1}^{n} \frac{r_{ij}^2}{\xi_i^2}$$

Or

$$\eta_j^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ij}^2}{\xi_i^2}$$

$$\xi_i^2 = \frac{1}{m} \sum_{j=1}^{m} \frac{r_{ij}^2}{\eta_j^2}$$

ADF

5 The Matrix Variate Normal

Gupta and Nagar (2000) chapter 2

If $\underline{x}_i$ are n iid normal $m-dimensional$ rv's with $\underline{x}_i \sim \mathcal{N}(\mu, \Sigma)$, then $\underline{X} \sim \mathcal{N}_{n,m}(I \otimes \mu, I \otimes \Sigma)$.

$$\underline{X} \sim \mathcal{N}(I \otimes \mu, I \otimes (AA' + \Delta^2))$$
$$\underline{X} = e\mu' + \underline{Z}A' + \underline{E}$$

References

Anderson, T. W., and H. Rubin. 1956. "Statistical Inference in Factor Analysis." In *Proeedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman, V:111–50. Berkeley; Los Angeles: University of California Press.

Bargmann, R. E. 1957. "A Study in Independence and Dependence in Multivariate Normal Analysis." Mimeograph Series 186. Institute of Statistics UNC. https://www.google.com/books/edition/A_Study_of_Independence_and_Dependence_i/tD1LAQAAMAAJ?hl=en.

Burt, C. 1947. "A Comparison of Factor Analysis and Analysis of Variance." *British Journal of Statistical Psychology* 1 (1): 3–26.

De Leeuw, J. 1973. "A Generalization of the Young-Whittle Model." Research Report 006-73. Leiden, The Netherlands: Department of Data Theory FSW/RUL.

Eckart, C., and G. Young. 1936. "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika* 1 (3): 211–18.

Gollob, H. F. 1968. "A Statistical Model Which Combines Features of Factir Analytic and Analysis of Variance Techniques." *Psychometrika* 33 (1): 73–115.

Gupta, A. K., and D. K. Nagar. 2000. *Matrix Variate Distributions*. Chapman & Hall.

Hemelrijk, J. 1966. "Underlining Random Variables." *Statistica Neerlandica* 20: 1–7.

Howe, W. G. 1955. "Some Contributions to Factor Analysis." ORNL 1919. Oak Ridge National Laboratory. https://technicalreports.ornl.gov/1955/3445603609833.pdf.

Jöreskog, K. G. 1962. "On the Statistical Treatment of Residuals in Factor Analysis." *Psychometrika* 27 (4): 335–54.

Lawley, D. N. 1940. "The Estimation of Factor Loadings by the Method of Maximum Likelihood." *Proceedings of the Royal Society of Edinburgh* A 60: 64–82.

———. 1941. "Further Investigations in Factor Estimation." *Proceedings of the Royal Society of Edinburgh* A 61: 176–85.

McDonald, R. P. 1979. "The Simultaneous Estimation of Factor Loadings and Scores." *British Journal of Mathematical and Statistical Psychology* 32 (212–228).

Whittle, P. 1952. "On Principal Components and Least Squares Methods of Factor Analysis." *Scandinavian Actuarial Journal* 3-4: 223–39.

Young, G. 1940. "Maximum Likelihood and Factor Analysis." *Psychometrika* 6 (1): 49–53.

Young, G., and A. S. Householder. 1938. "Discussion of a Set of Points in Terms of Their Mutual Distances." *Psychometrika* 3 (19-22).