Robust Factor Analysis

Peter Filzmoser

Dept. of Statistics, Vienna University of Technology
Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
Email: P.Filzmoser@tuwien.ac.at

Abstract: Two robust approaches to factor analysis are presented and compared. The first one uses a robust covariance matrix for estimating the factor loadings and the specific variances. The second one estimates factor loadings, scores and specific variances directly, using the alternating regression technique.

Keywords: Factor analysis, Alternating regression, Outliers, Robustness.

1. The Factor Analysis (FA) Model

Factor analysis (FA) is a standard technique in multivariate analysis which is routinely used in social and behavioral sciences. The aim of factor analysis is to summarize the correlation structure of observed variables $X_1, X_2, \ldots, X_p$. For this purpose one constructs $k < p$ unobservable or latent variables $f_1, \ldots, f_k$, which are called the factors, and which are linked with the original variables through the equation

$$X_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \ldots + \lambda_{jk} f_k + \varepsilon_j \tag{1}$$

for each $1 \leq j \leq p$. The error variables $\varepsilon_1, \ldots, \varepsilon_p$ are supposed to be independent, but they have *specific variances* $\psi_1, \ldots, \psi_p$. The coefficients λ_{jl} are called factor *loadings*, and they are collected into the matrix of loadings $\mathbf{\Lambda}$.

Using the vector notations $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$, $\boldsymbol{F} = (f_1, \ldots, f_k)^\top$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^\top$, the usual conditions on factors and error terms can be written as $E(\boldsymbol{F}) = E(\boldsymbol{\varepsilon}) = 0$, $\mathrm{Cov}(\boldsymbol{F}) = \boldsymbol{I}_k$, and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$, with $\boldsymbol{\Psi}$ a diagonal matrix containing on its diagonal the specific variances. Furthermore, $\boldsymbol{\varepsilon}$ and $\boldsymbol{F}$ are assumed to be independent.

In FA, one needs to estimate the matrix $\mathbf{\Lambda}$ (which is only specified up to an orthogonal transformation) and $\boldsymbol{\Psi}$. Classical FA methods are however very vulnerable to the presence of outliers, hence methods need to be constructed which can resist the effect of outliers.

2. FA using Robust Covariance Matrices

Denote by $\mathbf{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$ the covariance matrix of $\boldsymbol{X}$. (In case that $X_1, \ldots, X_p$ are standardized versions of the originally measured variables, $\mathbf{\Sigma}$ becomes the correlation matrix.) It follows from (1) that $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \boldsymbol{\Psi}$. In classical FA, the matrix $\mathbf{\Sigma}$ is estimated by the sample covariance matrix, which is afterwards decomposed to obtain the estimators for $\mathbf{\Lambda}$ and $\boldsymbol{\Psi}$. Many methods have been proposed for this decomposition, of which maximum likelihood (ML) and the principal factor analysis (PFA) method are the most frequently used. It is, however, well known that outliers can heavily influence the estimation of $\mathbf{\Sigma}$ and hence also the parameter estimates. Therefore a robust scatter matrix estimator needs to be used.

For this, it is convenient to use the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985). The MCD looks for the subset of h out of all n observations having the smallest determinant of its covariance matrix (typically, $h \approx 3n/4$). The MCD estimator is highly robust, has good efficiency properties and is available in several software packages. Recently, a fast MCD algorithm has been developed (Rousseeuw and Van Driessen, 1999).

Simulations and examples have shown that PFA based on MCD results in a resistant FA-method, with bounded influence function (Pison et al., 2002). The empirical influence function can be used as a data-analytic tool.

3. FA using Robust Alternating Regressions

A limitation of the MCD-based approach is that the sample size n needs to be bigger than the number of variables p. For samples with $n \leq p$ (which occur quite frequently in practice), the technique of alternating regressions can be used. For this we consider herefore the sample version of model (1):

$$X_{ij} = \sum_{l=1}^{k} \lambda_{jl} f_{il} + \varepsilon_{ij} \qquad (2)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Suppose that preliminary estimates for the *factor scores* f_{il} are known, and consider them as constants for a moment. The loadings λ_{jl} can now be estimated by linear regressions of the X_j's on the factors. Moreover, by applying a robust scale estimator on the computed residuals, estimates $\hat{\psi}_j$ for ψ_j can easily be obtained. On the other hand, if we take i fixed in (2) and suppose that the λ_{jl} are fixed, a regression of X_{ij} on the loadings λ_{jl} yields updated estimates for the factor scores. Since there is heteroscedasticity, weights proportional to $(\hat{\psi}_j)^{-1/2}$ should be included.

Using appropriate starting values for the factor scores, an iterative process (called alternating or criss-cross regressions) can be carried out to estimate the unknown parameters of the factor model. To make things robust, we propose to use robust regression procedures. Our suggestion is to use a weighted L_1-regression estimator since it is fast to compute and very robust (compare Croux et al., 2002). Experiments on real and simulated data show that this method works well, converges quite fast and is highly robust. A documented S-plus program computing the robust alternating regression estimator is freely available at *http://www.statistik.tuwien.ac.at/public/filz/* .

References:

C. Croux, P. Filzmoser, G. Pison, and P.J. Rousseeuw. Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 2002. To appear.

G. Pison, P.J. Rousseeuw, P. Filzmoser, and C. Croux. Robust factor analysis. *Journal of Multivariate Analysis*, 2002. To appear.

P.J. Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications, Vol. B* (eds. W. Grossmann et al.), pp. 283–297. Dordrecht: Reidel Publishing Co, 1985.

P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212-223, 1999.