A Robust Version of Principal Factor Analysis

G. Pison[1], P. J. Rousseeuw[1], P. Filzmoser[2] & C. Croux[3]

[1] Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Antwerpen, Belgium.
[2] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria.
[3] ECARE and Institut de Statistique, Université Libre de Bruxelles, CP-139, Avenue Roosevelt 50, B-1050 Bruxelles, Belgium.

Abstract. Our aim is to construct a factor analysis method that can resist the effect of outliers. We start with a highly robust initial covariance estimator, after which the factors can be obtained from maximum likelihood or from principal factor analysis (PFA). We find that PFA based on the minimum covariance determinant scatter matrix works well. We also derive the influence function of the PFA method. A new type of empirical influence function (EIF) which is very effective for detecting influential data is constructed. If the data set contains fewer cases than variables, we estimate the factor loadings and scores by a robust interlocking regression algorithm.

Keywords. Factor Analysis, Influence Function, Outliers, Robustness

1 The Factor Analysis Model

One aim of a factor analysis (FA) is to summarize the correlation structure of some observed variables $X_1, X_2, \ldots, X_p$. For this purpose it introduces $k < p$ unobservable or latent variables $\Phi_1, \ldots, \Phi_k$ which are called *factors*, and which are linked with the original variables through the equation

$$X_j = \mu_j + \Lambda_{j1}\Phi_1 + \Lambda_{j2}\Phi_2 + \ldots + \Lambda_{jk}\Phi_k + \varepsilon_j \tag{1}$$

for each $1 \leq j \leq p$. The error variables $\varepsilon_1, \ldots, \varepsilon_p$ are assumed to be independent, with $\varepsilon_j \sim N(0, \psi_j)$ where $\psi_1, \ldots, \psi_p$ are called the *specific variances*. The coefficients Λ_{jl} are called the factor *loadings*, and they are collected into the matrix of loadings $\boldsymbol{\Lambda}$.

Using the vector notations $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_k)^t$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^t$, the usual conditions on factors and error terms can be written as $E(\boldsymbol{\Phi}) = E(\boldsymbol{\varepsilon}) = 0$, $\mathrm{Cov}(\boldsymbol{\Phi}) = I_p$, and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$, with $\boldsymbol{\Psi}$ a diagonal matrix containing on its diagonal the specific variances. Furthermore, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\Phi}$ are assumed to be independent.

In factor analysis, one needs to estimate the matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ (the latter is only specified up to an orthogonal transformation). Let us denote the covariance matrix of X as $\boldsymbol{\Sigma}$. (In case that $X_1, \ldots, X_p$ are standardized versions of the originally measured variables, $\boldsymbol{\Sigma}$ represents a correlation matrix.) It follows from (1) that

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^t + \boldsymbol{\Psi} \tag{2}$$

under the model assumptions. In classical factor analysis the matrix $\boldsymbol{\Sigma}$ is estimated by the sample covariance matrix $\boldsymbol{S}$. Afterwards one tries to decompose $\boldsymbol{S}$ as in (2) to obtain the estimators for $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. Typically $\boldsymbol{S}$ cannot

be decomposed exactly as in (2), so we must resort to an approximate decomposition. Many methods have been proposed for this, of which maximum likelihood estimation (MLE) and the principal factor analysis (PFA) method are the most frequently used (see, e.g. Basilevsky 1994).

It is, however, well known that outliers can heavily influence the estimation of $\boldsymbol{\Sigma}$ and hence also the parameter estimates. Our aim is to construct a factor analysis method that can resist the effect of outliers. Therefore a robust covariance matrix estimator needs to be used. It is convenient to use the Minimum Covariance Determinant (MCD) estimator (Rousseeuw 1985). The MCD looks for the subset of h observations having the smallest determinant of its covariance matrix. Typically, $h \approx 3n/4$. The MCD estimator of $\boldsymbol{\Sigma}$ is then the covariance matrix computed from that subset, and the MCD estimator of $\boldsymbol{\mu}$ is its mean. The MCD estimator is highly robust and has good efficiency properties (Croux and Haesbroeck, 1999). Moreover, a fast algorithm for the MCD has recently been developed (Rousseeuw and Van Driessen, 1999).

The resulting robust loadings $\boldsymbol{L}_n^r$ and specific variances $\boldsymbol{P}_n^r$ allow us to obtain robust factors $\boldsymbol{F}_n^r$. They describe the correlation or covariance between the uncontaminated data.

2 Empirical Study

Some empirical studies with outliers are carried out to investigate their effect on classical and robust FA.

2.1 Sensitivity Analysis

We investigate the sensitivity of factor analysis to outliers and small errors. Therefore we compare the sensitivity of classical maximum likelihood estimation (CLAS.MLE), principal factor analysis (CLAS.PFA), and their MCD-based versions on the stock price data set of (Johnson and Wichern 1998), with $n = 100$ observations and $p = 5$ variables. The data are standardized by subtracting the average of each variable and dividing by its standard deviation.

We first estimate $k = 2$ factors based on the classical and robust correlation matrices, yielding the loadings $\boldsymbol{L}_n^{(0)} \in I\!R^{5\times 2}$ and unique variances $\boldsymbol{P}_n^{(0)} = (P_1^{(0)}, \ldots, P_5^{(0)})$. For the sensitivity analysis we add a noise matrix $(err^{(s)})$ and a matrix $(xout^{(s)})$ which causes n_{out} data points to become outliers. The elements of the noise matrix are distributed according to $N(0, (0.05)^2)$. The outlier matrix $xout^{(s)}$ is mainly zero, except for n_{out} elements. We generate only one outlying entry per outlying object, which is generated from the normal distribution $N(10, (0.05)^2)$. The disturbed data sets $\boldsymbol{X}^{(s)}$ are thus generated as $\boldsymbol{X}^{(s)} = \boldsymbol{X}^{(0)} + err^{(s)} + xout^{(s)}$ for $s = 1, \ldots, m$. Fitting this model yields estimates $\boldsymbol{L}_n^{(s)}$ and $\boldsymbol{P}_n^{(s)}$ for $m = 1000$ simulated samples. Since the loadings matrix is only determined up to an orthogonal matrix, we consider the $p \times p$ matrix $\boldsymbol{A}_n^{(s)} = \boldsymbol{L}_n^{(s)}(\boldsymbol{L}_n^{(s)})^t$ instead. We will compare the elements $a_{ij}^{(s)}$ of $\boldsymbol{A}_n^{(s)}$ with the undisturbed entries $a_{ij}^{(0)}$ of the matrix $\boldsymbol{A}_n^{(0)} = \boldsymbol{L}_n^{(0)}(\boldsymbol{L}_n^{(0)})^t$ by computing the mean squared error (MSE) of the estimates:

$$MSE(a_{ij}) = \frac{1}{m}\sum_{s=1}^{m}\left(a_{ij}^{(s)} - a_{ij}^{(0)}\right)^2$$

for $i, j = 1, \ldots, p$. The average MSE is $MSE(\boldsymbol{A}) = \frac{1}{p^2} \sum_{i=1}^{p} \sum_{j=1}^{p} MSE(a_{ij})$. Similarly, for the square root of the unique variances P_j we compute

$$MSE(P_j) := \frac{1}{m} \sum_{s=1}^{m} \left(\sqrt{P_j^{(s)}} - \sqrt{P_j^{(0)}} \right)^2$$

where $j = 1, \ldots, p$ and the average MSE is $MSE(\boldsymbol{P}) = \frac{1}{p} \sum_{j=1}^{p} MSE(P_j)$.

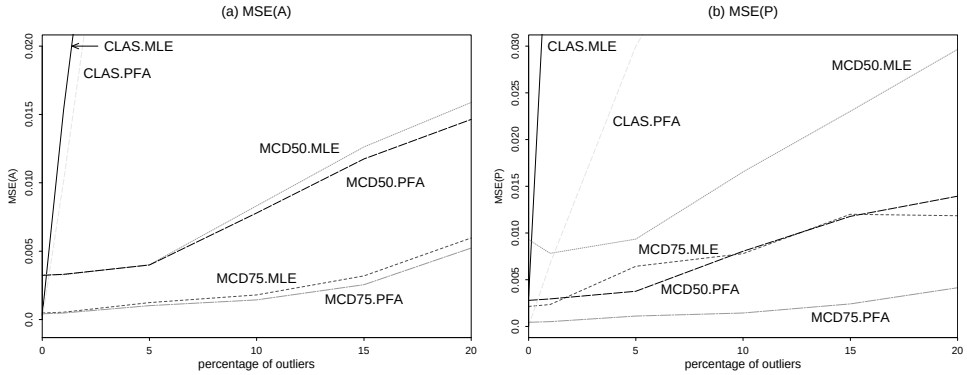

Fig. 1. Sensitivity of factor analysis on the stock price data: (a) $MSE(\boldsymbol{A})$ versus the fraction of outliers; (b) $MSE(\boldsymbol{P})$ versus the fraction of outliers.

Figure 1 shows the average MSE versus the fraction of outliers (here, 0% to 20%). We see that the MSE's of CLAS.PFA and CLAS.MLE are much higher than those based on the robust correlation matrix. Comparing MCD50 ($h \approx 0.5n$) and MCD75 ($h \approx 0.75n$), we find that FA using MCD75 systematically yielded a lower MSE than the corresponding method based on MCD50. Because MCD75 also has a higher efficiency than MCD50, we will work with MCD75 from now on.

2.2 Monte Carlo study

We start from fixed parameter values, i.e. $\boldsymbol{\Lambda}$ and a diagonal matrix $\boldsymbol{\Psi}$. Then we construct data sets $\boldsymbol{X}^{(s)}$ as follows: $\boldsymbol{X}^{(s)} = \boldsymbol{\Lambda}\boldsymbol{\Phi}^{(s)} + \boldsymbol{\varepsilon}^{(s)} + \text{Out}^{(s)}$. For each s we generated the matrix of factor scores $\boldsymbol{\Phi}^{(s)}$ from $N(0, 1)$, the entries $\varepsilon_{ij}^{(s)}$ of the noise term $\boldsymbol{\varepsilon}^{(s)}$ from $N(0, \psi_j)$, and the outlying term $\text{Out}^{(s)}$ from $N(10, (0.05)^2)$.

Fitting the factor analysis model to $\boldsymbol{X}^{(s)}$ gives the estimates $\boldsymbol{L}_n^{(s)}$ and $\boldsymbol{P}_n^{(s)}$ for $s = 1, \ldots, 1000$. These estimates are compared to the true $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ by computing the MSE.

From the simulations (with $n = 100$, $p = 5$, and $k = 2$) we conclude that PFA based on MCD is more robust than MLE based on MCD. Therefore, we will focus on the MCD75.PFA technique.

3 The Influence Function of PFA

3.1 The Theoretical Influence Function

We have derived the influence functions $IF(x, \boldsymbol{P}, G)$ and $IF(x, \boldsymbol{L}\boldsymbol{L}^t, G)$ of the PFA method. In Figure 2 we plot the influence function of the first specific variance.

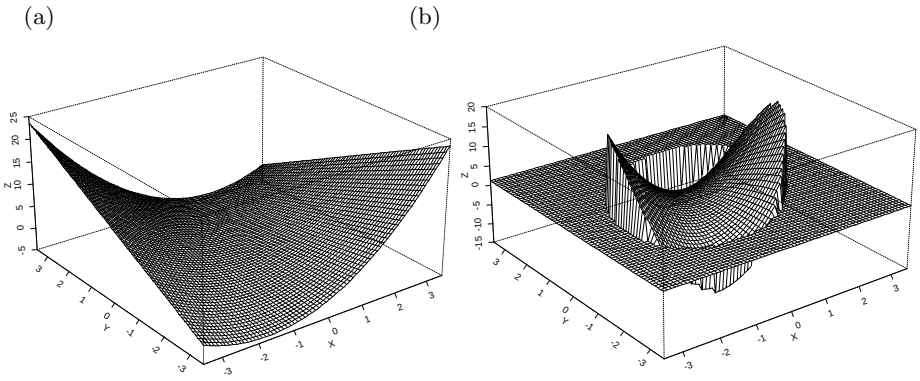

Fig. 2. Influence function $IF(x, P_1, G)$ based on (a) the classical covariance matrix and (b) the MCD75 scatter matrix.

We see that for the PFA based on the classical covariance matrix, the influence functions of ψ_j and $(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^t)_{ij}$ are unbounded. This shows that an outlying x can have an arbitrarily large effect on the PFA method. On the other hand the influence function of MCD-based PFA is bounded, so an outlier x has a bounded effect on the robust PFA. Inside the elliptical central region of the x-distribution the IF looks like that of the classical PFA. Outside, the influence function is a constant for the MCD covariance matrix, and it becomes zero when working with the MCD correlation matrix.

3.2 The Empirical Influence Function

In the previous subsection, we computed the influence functions in the population case, where we know the true underlying distribution G. In the empirical setting we only have a sample $\boldsymbol{X}_n \in I\!R^{n \times p}$ without knowing G. However, the unknown G depends only on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which we will replace by estimates $\boldsymbol{T}(\boldsymbol{X}_n)$ and $\boldsymbol{S}(\boldsymbol{X}_n)$ in the formula of the influence function. The resulting *empirical influence function* (EIF) is then evaluated in each data point x_i $i = 1, \ldots, n$ to measure their effect on the principal factor analysis. Our aim is to detect the most influential observations x_i by comparing the EIF(x_i).

We can construct the EIF of the classical PFA (e.g. of $\boldsymbol{P}_n^c$) and of the robust PFA (e.g. of $\boldsymbol{P}_n^r$). For $\boldsymbol{T}(\boldsymbol{X}_n)$ and $\boldsymbol{S}(\boldsymbol{X}_n)$ we take the classical estimates $(\boldsymbol{T}_n^c, \boldsymbol{S}_n^c)$ or the robust estimates $(\boldsymbol{T}_n^r, \boldsymbol{S}_n^r)$. This yields three ways to define the EIF:

- Tanaka and Odaka (1989) computed $EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^c, \boldsymbol{S}_n^c)$. This approach often masks outliers because $\boldsymbol{T}_n^c$ and $\boldsymbol{S}_n^c$ break down.
- Substituting the robust $\boldsymbol{T}_n^r$ and $\boldsymbol{S}_n^r$ in the robust IF yields $EIF(x_i; \boldsymbol{P}_n^r; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)$. This function illustrates the fact that an outlying x_i has only a small effect on $\boldsymbol{P}_n^r$, which is natural because we constructed $\boldsymbol{P}_n^r$ for this purpose.
- Substituting the robust $\boldsymbol{T}_n^r$ and $\boldsymbol{S}_n^r$ in the classical IF yields $EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)$. This is the most useful, because $\boldsymbol{T}_n^r$ and $\boldsymbol{S}_n^r$ are not affected by outliers. Ideally, we would like to have $EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the true $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the parent distribution, but in the presence of outliers the $\boldsymbol{T}_n^r$ and $\boldsymbol{S}_n^r$ are good approximations to these parameters.

In practice, to detect the most influential data points x_i (i.e. points that would strongly affect the classical PFA) we therefore recommend to compute the $EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)$.

4 Example

The aircraft data set (Gray 1985) consists of $n = 23$ single-engine aircraft built in the years 1947-1979 with $p = 5$ variables. Applying the MCD to these data indicates that cases 14 and 22 are outliers. Let us now compute the empirical influence functions $EIF(x_i; P_j)$ and an overall value $\|EIF(x_i; \boldsymbol{P})\| = \sqrt{|EIF(x_i; P_1)|^2 + \ldots + |EIF(x_i; P_5)|^2}$ in the 23 observations x_i for the different versions of the EIF considered above. Figure 3 plots $\|EIF(x_i; \boldsymbol{P})\|$ versus the case number i. We see that the outlying cases 14 and 22 have a

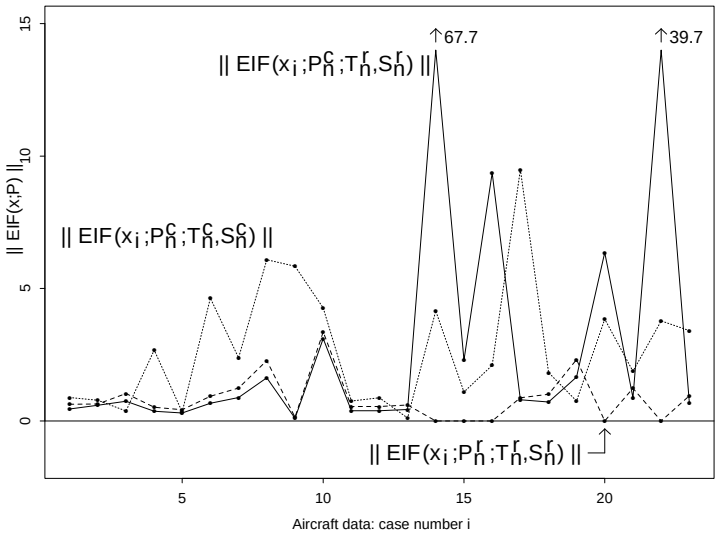

Fig. 3. The empirical influence functions $\|EIF(x_i; \boldsymbol{P})\|$ evaluated in 23 aircraft.

relatively small $\|EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^c, \boldsymbol{S}_n^c)\|$. This is because $\boldsymbol{T}_n^c$ and $\boldsymbol{S}_n^c$ try to fit all the data points, so $\boldsymbol{S}_n^c$ becomes too large. Secondly, using the robust estimates $\boldsymbol{P}_n^r$, $\boldsymbol{T}_n^r$ and $\boldsymbol{S}_n^r$ leads to $\|EIF(x_i; \boldsymbol{P}_n^r; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)\| = 0$ for cases 14

and 22. This illustrates the robustness of $\boldsymbol{P}_n^r$ but does not help to detect the influential points. The only function that clearly shows the influential points is $\|EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)\|$, which takes on huge values for cases 14 and 22. In this example the outliers for $\boldsymbol{S}_n^r$ are also the most influential points for PFA. This does not always have to be the case.

5 Robust Interlocking Regression

When the data set contains fewer cases than variables, we can not compute the MCD covariance matrix. In this case an approach called "Factor Analysis using Interlocking Regressions" (FAIR) can be used. This method estimates the unknown parameters (factor scores and loadings) directly, without passing via an estimate of the covariance matrix.

Let the variables already be standardized to have zero location and unit spread. We denote the vector of unknown parameters as $\boldsymbol{\theta} = (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_n, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_p)$. The FAIR estimator is defined by:

$$\hat{\boldsymbol{\theta}}_{FAIR} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{p} w_i(\boldsymbol{\theta}) v_j(\boldsymbol{\theta}) |x_{ij} - \hat{x}_{ij}(\boldsymbol{\theta})|$$

with $w_i(\boldsymbol{\theta}) = \min(1, \chi^2_{k,0.95}/\mathrm{RD}_i^2)$ for $i = 1, \ldots, n$ where the RD_i are robust distances computed from the collection of score vectors. Analogously, the set of column weights v_j is defined using the loading vectors. The FAIR estimator is extensively discussed in Croux, et al (1999). It can cope with many outlying cells in the data matrix, without loosing too much efficiency. A robust interlocking regression algorithm is available to compute the FAIR estimator.

References

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York: Wiley.

Croux, C., Filzmoser, P., Pison, G. and Rousseeuw, P.J. (1999). Fitting Factor Models by Robust Interlocking Regression. Submitted for publication.

Croux, C. and Haesbroeck, G. (1999). Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis*, **71**, 161–190.

Gray, J.B. (1985). Graphics for Regression Diagnostics. In: *American Statistical Association Proceedings of the Statistical Computing Section*, 102–107. Washington, D.C.: ASA.

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point. In: *Mathematical Statistics and Applications, Vol. B*, 283-297. Dordrecht: Reidel.

Rousseeuw, P.J. and Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**, 212-223.

Tanaka, Y. and Odaka, Y. (1989). Influential Observations in Principal Factor Analysis. *Psychometrika*, **54**, 475–485.