Robust Factor Analysis

Greet Pison, Peter J. Rousseeuw,
Peter Filzmoser, and Christophe Croux

August 1, 2001

Abstract

Our aim is to construct a factor analysis method that can resist the effect of outliers. For this 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 based on either the classical scatter matrix or a robust matrix. These results are applied to the construction of a new type of empirical influence function (EIF) which is very effective for detecting influential data. To facilitate the interpretation, we compute a cutoff value for this EIF. Our findings are illustrated with several real data examples.

Keywords: Factor Analysis, Influence Function, Multivariate Analysis, Outlier Detection, Robust Estimation.

Greet Pison is Assistant and Peter J. Rousseeuw is Professor, Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium. Peter Filzmoser is Assistant, Department of Statistics, Probability Theory and Actuarial Mathematics, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria. Christophe Croux is Professor, ECARES and Institut de Statistique, Université Libre de Bruxelles, CP-114, Avenue Roosevelt 50, B-1050 Brussels, Belgium. We are grateful to the late Bernhard Flury for providing us with the data sets in Section 5.

1 Introduction

Factor analysis is a popular multivariate technique. Its goal is to approximate the p original variables of the dataset by linear combinations of a smaller number k of latent variables, called *factors*. This must be done in such a way that the covariance matrix (or the correlation matrix) of the p original variables is fitted well. The factor analysis model contains many parameters, including the specific variances of the error components.

The assumptions underlying the factor analysis model are rather strong compared to its applications. Therefore many authors have investigated whether these assumptions are necessary. It was already shown that the classical estimates have good asymptotic properties under some weaker assumptions (see e.g. Browne and Shapiro 1988, Mooijaart and Bentler 1991).

The classical technique starts by computing the usual sample covariance matrix or the sample correlation matrix, followed by a second step which decomposes this matrix according to the model. This approach is not robust to outliers in the data, since they already have a large effect on the first step. In Section 2 we therefore construct a robust factor analysis method, which in the first step computes a highly resistant scatter matrix such as the minimum covariance determinant (MCD) estimator (Rousseeuw 1985). In the context of structural equation models, Yuan and Bentler (1998a, 1998b) used M-estimators (Maronna 1976) and S-estimators (Davies 1987, Rousseeuw and Leroy 1987) and minimized the resulting Wishart likelihood function. For the second step several methods are available, such as maximum likelihood estimation and the principal factor analysis method (PFA). The simulations in Section 3 yield a slight preference for the latter.

In order to study the robustness of the PFA method we compute its influence function (the complete derivation can be found in the Appendix). The influence function depends, among other things, on the scatter matrix estimator of the first step. Section 4 plots the influence function of PFA based on the classical covariance matrix and compares it with that based on the MCD. The latter influence function is bounded. We also study the influence function of PFA applied to the robust correlation matrix derived from the MCD, and find that the influence of a far outlier becomes exactly zero.

Not all outliers have a large influence on the factor analysis. In order to detect influential data points we construct an empirical influence function (EIF) in Section 4.2. We argue that the most informative version is the EIF of the *classical* PFA, but evaluated in the distribution

characterized by the *robust* estimates of location and scatter. Moreover, we compute a cutoff value for the EIF to tell us when a data point is truly influential. Section 5 illustrates the robust approach on two real data examples.

2 The Factor Analysis Model

Classical factor analysis tries to describe the correlation matrix $\boldsymbol{\rho}$ or the covariance matrix $\boldsymbol{\Sigma}$ between the original variables $X_1, X_2, \ldots, X_p$ by a small number $k \leq p$ of new variables $\Phi_1, \ldots, \Phi_k$ called *factors*. These factors are unobservable. In particular, the orthogonal factor analysis model says that

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\boldsymbol{\Phi} + \boldsymbol{\varepsilon} \tag{2.1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^t$ is the mean vector, $\boldsymbol{\Lambda} \in I\!R^{p \times k}$ is the matrix of factor loadings, $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_k)^t$, and the error term is $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^t$. Note that the matrix $\boldsymbol{\Lambda}$ is only determined up to right multiplication by an orthogonal matrix $\boldsymbol{U}$. We assume that the random vectors $\boldsymbol{\Phi}$ and $\boldsymbol{\varepsilon}$ are independent, $E(\boldsymbol{\Phi}) = \boldsymbol{0}$, $Cov(\boldsymbol{\Phi}) = \boldsymbol{I}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $Cov(\boldsymbol{\varepsilon}) = diag(\boldsymbol{\Psi})$ with $\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_p) \in I\!R^p$. Under these assumptions we obtain

$$\boldsymbol{\Sigma} \;=\; \boldsymbol{\Lambda}\boldsymbol{\Lambda}^t + diag(\boldsymbol{\Psi}). \tag{2.2}$$

Because of the number of parameters in this model, for a given p the largest possible k is

$$[p + 0.5 - \sqrt{2p + 0.25}]$$

(see, e.g., Johnson and Wichern 1998, page 538) where $[\ldots]$ stands for the integer part of a real number. For instance, for 5-variate $\boldsymbol{X}$ we can estimate up to 2 factors.

In practice, we have a data set with n objects in p dimensions. The classical factor analysis method computes the sample mean vector $\boldsymbol{T}_n^c$ to estimate $\boldsymbol{\mu}$, and the sample covariance matrix $\boldsymbol{S}_n^c$ to estimate $\boldsymbol{\Sigma}$. (Throughout, the superscript c stands for *classical*, i.e. based on Gaussian distributions.) Afterwards a decomposition like (2.2) is carried out to obtain an estimate $\boldsymbol{L}_n$ for $\boldsymbol{\Lambda}$ and an estimate $\boldsymbol{P}_n$ for $\boldsymbol{\Psi}$, thereby yielding an estimate $\boldsymbol{F}_n$ for $\boldsymbol{\Phi}$. Many methods have been proposed for this decomposition, of which the maximum likelihood estimator (MLE) and the principal factor analysis (PFA) algorithms are the most frequently

used (see e.g., Basilevsky 1994). The MLE method minimizes the log-likelihood function

$$\mathcal{L}(\mathbf{\Lambda}, \mathbf{\Psi}) = c[-ln|\mathbf{\Lambda}\mathbf{\Lambda}^t + diag(\mathbf{\Psi})| + tr[\hat{\boldsymbol{S}}(\mathbf{\Lambda}\mathbf{\Lambda}^t + diag(\mathbf{\Psi}))^{-1}]$$

with c some constant (see Jöreskog 1963). For $\hat{\boldsymbol{S}}$ we can use $\boldsymbol{S}_n^c$ in the classical case and $\boldsymbol{S}_n^r$ in the robust method. The principal factor analysis is based on eigenvalue/eigenvector analysis of the reduced covariance matrix, so here again we use $\boldsymbol{S}_n^c$ in the classical case and $\boldsymbol{S}_n^r$ in the robust method.

Since these methods cannot resist the effect of outliers, we propose to start from a more robust location vector and scatter matrix. It is convenient to use the Minimum Covariance Determinant Estimator (MCD) of Rousseeuw (1984, 1985). The MCD looks for that h-subset of the data with the smallest determinant of its covariance matrix. Typically, $h \approx 3n/4$. The MCD location $\boldsymbol{T}_n^r$ is then the average of the h points in that subset, and the MCD scatter estimate $\boldsymbol{S}_n^r$ is a multiple of their covariance matrix. (Throughout, the superscript r stands for *robust*.) The MCD is highly robust and converges at a faster rate than the previously popular Minimum Volume Ellipsoid (MVE) estimator. Moreover, the MCD can now be computed very quickly with the new algorithm of Rousseeuw and Van Driessen (1999).

The resulting robust loadings $\boldsymbol{L}_n^r$ and specific variances $\boldsymbol{P}_n^r$ will be different from the classical $\boldsymbol{L}_n^c$ and $\boldsymbol{P}_n^c$. Because the classical scatter matrix $\boldsymbol{S}_n^c$ is influenced by outlying data points, this is also the case for the resulting loadings $\boldsymbol{L}_n^c$, the specific variances $\boldsymbol{P}_n^c$ and the factor scores $\boldsymbol{F}_n^c$. On the other hand, the MCD scatter matrix is robust to outliers, so it allows us to obtain robust factors $\boldsymbol{F}_n^r$ which describe the correlation or covariance between the uncontaminated data. Let us look at a first example to illustrate this.

Example 1. The aircraft data set (Gray 1985) consists of $n = 23$ single-engine aircraft built in the years 1947-1979. The $p = 5$ variables are the aspect ratio, lift-to-drag ratio, weight of the plane, maximal thrust, and cost. Applying the MCD to these data indicates that cases 14 and 22 are outliers. Plane 22 was the F-111 aircraft. It was built to suit the needs of the Army, the Navy and the Air Force simultaneously. At that time, it was the most sophisticated, fastest, heaviest and most costly single-engine jet plane ever built. Nevertheless it had many technical problems. Plane 14 was the F-104A 'Starfighter' which had a huge lift-to-drag ratio.

Let us now estimate $k = 2$ factors. Applying the principal factor (PFA) method to the classical correlation matrix yields the biplot in Figure 1a. The biplot in Figure 1b was

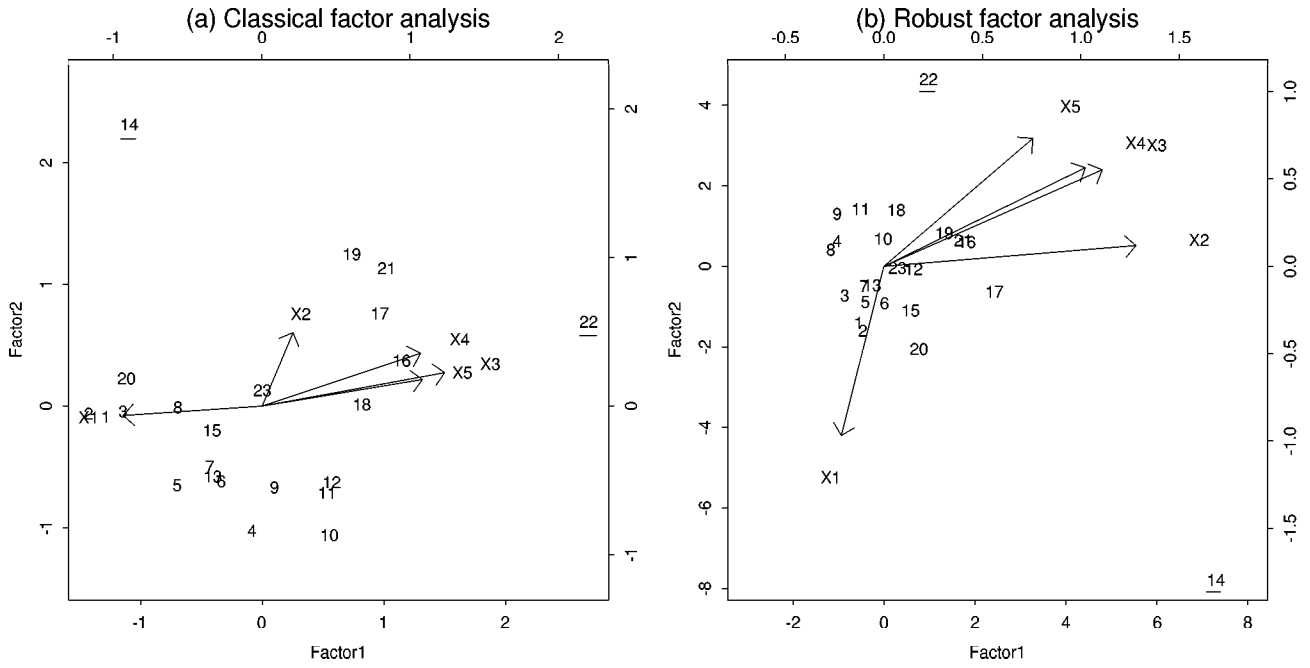

Figure 1: Biplots of (a) classical, and (b) robust factor analysis on the aircraft data set. The two outlying planes (14 and 22) are underlined.

obtained by applying PFA to the MCD-based robust correlation matrix $\boldsymbol{R}_n^r$ computed as

$$\boldsymbol{R}_n^r = \boldsymbol{D}\boldsymbol{S}_n^r\boldsymbol{D} \qquad \text{with} \qquad \boldsymbol{D} = \text{diag}(((\boldsymbol{S}_n^r)_{11})^{-1/2}, \ldots, ((\boldsymbol{S}_n^r)_{pp})^{-1/2}). \tag{2.3}$$

In the biplot (Gabriel 1971) the arrows indicate the positions of the variables by plotting (L_{j1}, L_{j2}) for $j = 1, \ldots, p$. The observations (F_{i1}, F_{i2}) are also added on the plot. The main idea is that the biplot represents the general interaction structure between the variables and the observations. More details on biplots can be found in Gower and Hand (1996).

The main difference between the two methods is that in the classical factor analysis the two outliers highly influence $\boldsymbol{S}_n^c, \boldsymbol{L}_n^c$, and $\boldsymbol{F}_n^c$. So, also the classical biplot was influenced by these outliers. The robust factor analysis downweights these outliers, and gives a more reliable picture of the majority of the data. In this case the robust biplot represents the structure of the good observations and therefore this biplot ressembles the usual biplot based on the clean data. Let us compare the loadings of the classical and the robust factor analysis in Table 1. In the classical case, factor 1 was mainly a combination of variables 1 (with negative coefficient), 3, 4, and 5, and factor 2 was mostly determined by variable 2. In the robust factor analysis, factor 1 is a positive combination of variables 2, 3, and 4, whereas factor 2 essentially combines variables 1 and 5 (with different signs). We also see

Table 1: Loadings of classical and robust factor analysis on the aircraft data set.

	Loadings of Classical FA		Loadings of Robust FA	
Variable	Factor 1	Factor 2	Factor 1	Factor 2
X1: Aspect Ratio	-0.710	0.000	-0.165	-0.898
X2: Lift-to-Drag	0.157	0.672	0.981	0.110
X3: Weight	0.932	0.306	0.849	0.513
X4: Thrust	0.807	0.485	0.783	0.523
X5: Cost	0.818	0.244	0.580	0.679

that the second picture in Figure 1 is not simply a rotation of the first. In this example, the two methods give a quite different result.

3 Empirical Study

In this section we carry out empirical studies with outliers, to investigate their effect on classical and robust factor analysis. First we carry out a sensitivity analysis, and then a Monte Carlo experiment.

3.1 Sensitivity Analysis

We investigate the sensitivity of factor analysis to outliers and small errors. We will 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 stock price data set $\boldsymbol{X}^{(0)}$ contains the weekly returns of five stocks listed on the New York Stock Exchange. 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. For this we randomly choose n_{out} different rows in $xout^{(s)}$, and for each such row we choose a random entry. In these n_{out} entries of $xout$ we put values 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. The method for estimating the factor model was of course the same for the contaminated data as for the original data.

The estimates from the disturbed and the original data are compared in the following way. 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. More precisely, we compare the elements $a_{ij}^{(s)}$ of $\boldsymbol{A}_n^{(s)}$ with the undisturbed entries $a_{ij}^{(0)}$ of the matrix $\boldsymbol{A}_n^{(0)} = \boldsymbol{L}_n^{(0)}(\boldsymbol{L}_n^{(0)})^t$. For this we compute the mean squared error (MSE), bias (BIAS), and variance (VAR) of the estimates as

$$\begin{aligned}
MSE(a_{ij}) &= \frac{1}{m}\sum_{s=1}^{m}\left(a_{ij}^{(s)} - a_{ij}^{(0)}\right)^2 \\
BIAS(a_{ij}) &= \frac{1}{m}\sum_{s=1}^{m}\left(a_{ij}^{(s)} - a_{ij}^{(0)}\right) \\
VAR(a_{ij}) &= \frac{1}{m}\sum_{s=1}^{m}\left(a_{ij}^{(s)} - \frac{1}{m}\sum_{t=1}^{m} a_{ij}^{(t)}\right)^2
\end{aligned}$$

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

$$\begin{aligned}
MSE(P_j) &:= \frac{1}{m}\sum_{s=1}^{m}\left(\sqrt{P_j^{(s)}} - \sqrt{P_j^{(0)}}\right)^2 \\
BIAS(P_j) &:= \frac{1}{m}\sum_{s=1}^{m}\left(\sqrt{P_j^{(s)}} - \sqrt{P_j^{(0)}}\right) \\
VAR(P_j) &:= \frac{1}{m}\sum_{s=1}^{m}\left(\sqrt{P_j^{(s)}} - \frac{1}{m}\sum_{t=1}^{m}\sqrt{P_j^{(t)}}\right)^2
\end{aligned}$$

where $j = 1,\ldots,p$ and the average MSE is given by $MSE(\boldsymbol{P}) = \frac{1}{p}\sum_{j=1}^{p} MSE(P_j)$. However, it is well-known that the MLE and PFA methods may sometimes produce a negative estimate

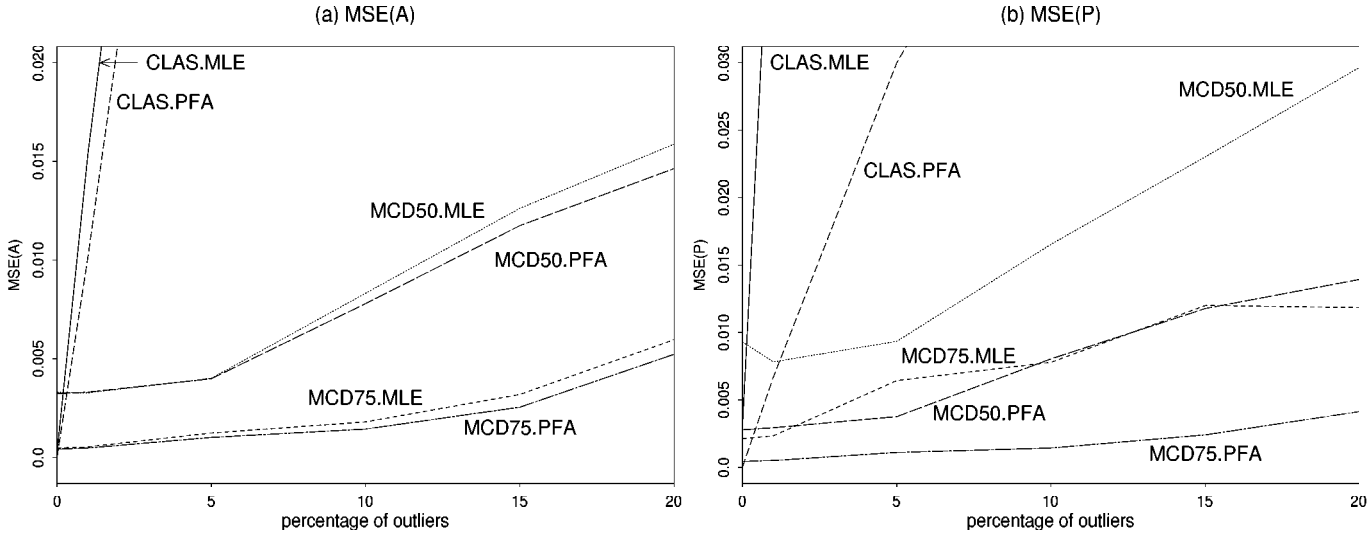

Figure 2: Sensitivity of factor analysis on the stock price data: (a) $MSE(\boldsymbol{A})$ versus the fraction of outliers; (b) $MSE(\boldsymbol{P})$ versus the fraction of outliers.

$P_j^{(s)}$. This is the so-called Heywood case (see Ten Berge and Kiers 1991, Kano 1998). In our simulation such negative $P_j^{(s)}$ occurred only a few times, with small values of $|P_j^{(s)}|$, so we have set these negative $P_j^{(s)}$ equal to zero.

For the stock price data, Figure 2 shows the average MSE versus the fraction of outliers (here, 0% to 20%). We can see that the MSE's of factor analysis based on the classical correlation matrix are much higher than those based on the robust correlation matrix using the MCD method. The fact that using a classical correlation matrix yields a higher MSE than using a more robust scatter matrix confirms the simulation of Kosfeld (1996) who inserted M-estimators of covariance. In Figure 2, MCD50 stands for the MCD estimator with $h \approx 0.5 * n$, and MCD75 corresponds to $h \approx 0.75 * n$. Comparing MCD50 and MCD75, we find that a factor analysis using MCD75 systematically yielded a lower MSE than the corresponding method based on MCD50. For other data sets, real and generated, we found similar results. Because MCD75 also has a higher efficiency than MCD50, we will work with MCD75 from now on.

3.2 Monte Carlo study

Here we do not start from a given data set but from fixed parameter values, i.e. an $n \times k$ matrix $\boldsymbol{\Lambda}$ and a $p \times p$ diagonal matrix $diag(\boldsymbol{\Psi})$. (The entries of $\boldsymbol{\Lambda}$ were generated from $N(0, \frac{1}{9})$ and those of $diag(\boldsymbol{\Psi})$ from the uniform distribution on the interval $[0, 1]$.) Then we

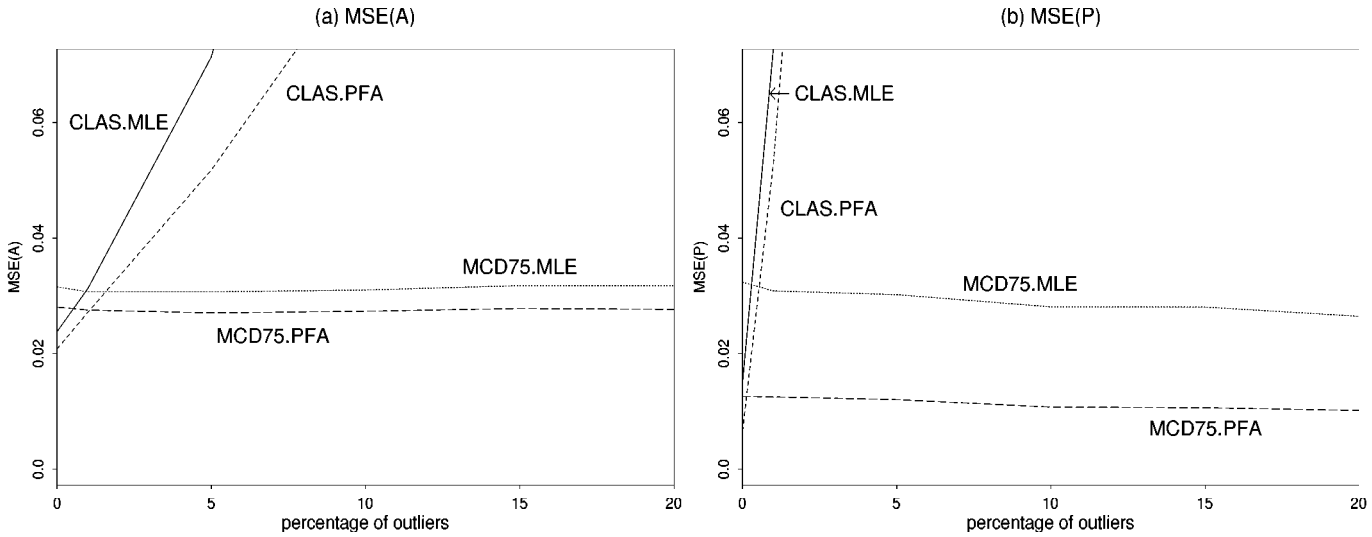

Figure 3: Simulation study: (a) $MSE(\boldsymbol{A})$ versus the fraction of outliers; (b) $MSE(\boldsymbol{P})$ versus the fraction of outliers.

construct data sets $\boldsymbol{X}^{(s)}$ according to the factor analysis model (2.1), i.e.

$$\boldsymbol{X}^{(s)} = \boldsymbol{\Lambda}\boldsymbol{\Phi}^{(s)} + \boldsymbol{\varepsilon}^{(s)} + \text{Out}^{(s)}.$$

For each s we generated the $k \times p$ matrix of factor scores $\boldsymbol{\Phi}^{(s)}$ from $N(0,1)$, and the entries $\varepsilon_{ij}^{(s)}$ of the noise term $\boldsymbol{\varepsilon}^{(s)}$ are distributed according to $N(0,\psi_j)$. The outlying term $\text{Out}^{(s)}$ was generated as in the previous subsection.

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

For the simulations in Figure 3 we took $n = 100$, $p = 5$, and $k = 2$. We see that the robust factor analysis based on MCD75 and the principal factor method gave the smallest mean squared error. Maximum likelihood estimation gave larger errors in all our simulations (also for other n and p). This parallels the results in Figure 2. Therefore, from now on we will focus on the MCD75.PFA technique.

4 The Influence Function of PFA

4.1 The Theoretical Influence Function

We now derive the theoretical influence function of the Principal Factor Analysis method. The influence function (see Hampel et al. 1986) of a functional Q at a distribution H

measures the effect on Q of adding a small mass at x. If we denote the point mass at x by $\triangle_x$ and write $H_t = (1-t)H + t\triangle_x$ then the influence function is given by

$$IF(x, Q, H) = \frac{\partial}{\partial t} Q(H_t)|_{t=0}.$$

In order to apply this we need the functional form of the PFA estimator. Let H be an arbitrary distribution on $\mathbb{R}^p$ with location estimate $\boldsymbol{T}(H) \in \mathbb{R}^p$ and scatter estimate $\boldsymbol{S}(H) \in \mathbb{R}^{p\times p}$. We will denote the PFA functional as $(\boldsymbol{A}(H), \boldsymbol{P}(H))$ where $\boldsymbol{A}(H) \in \mathbb{R}^{p\times p}$ is a positive semidefinite matrix with rank at most k, and $\boldsymbol{P}(H)$ is a vector in $\mathbb{R}^p$ with nonnegative components. The fitted scatter matrix is then

$$\boldsymbol{A}(H) + diag(\boldsymbol{P}(H)).$$

The PFA functional is defined as the pair $(\boldsymbol{A}(H), \boldsymbol{P}(H))$ that gives the closest fit to the observed $\boldsymbol{S}(H)$. Formally,

$$\begin{aligned}(\boldsymbol{A}(H), \boldsymbol{P}(H)) &= \underset{(\boldsymbol{A},\boldsymbol{P})}{\operatorname{argmin}} \sum_{i=1}^{p}\sum_{j=1}^{p}((\boldsymbol{S}(H))_{ij} - (\boldsymbol{A} + diag(\boldsymbol{P}))_{ij})^2 \\ &= \underset{(\boldsymbol{A},\boldsymbol{P})}{\operatorname{argmin}} \text{ trace } ((\boldsymbol{S}(H) - \boldsymbol{A} - diag(\boldsymbol{P}))(\boldsymbol{S}(H) - \boldsymbol{A} - diag(\boldsymbol{P}))^t). \qquad (4.1)\end{aligned}$$

So we use a least squares criterion to measure the closeness between $\boldsymbol{S}(H)$ and $\boldsymbol{A} + diag(\boldsymbol{P})$. Alternatively, one could use weighted least squares or a likelihood criterium here. Such an approach would of course yield an estimator different from the PFA-solution.

The spectral decomposition of $\boldsymbol{A}(H)$ yields

$$\boldsymbol{A}(H) = \sum_{j=1}^{k} \lambda_j(H)\boldsymbol{v}_j(H)\boldsymbol{v}_j(H)^t \qquad (4.2)$$

with eigenvalues $\lambda_j(H) > 0$ and orthonormal eigenvectors $\boldsymbol{v}_j(H)$ for $j = 1, \ldots, k$. Minimizing (4.1) yields two first order equations:

$$(\boldsymbol{S}(H) - diag(\boldsymbol{P}(H)))\boldsymbol{v}_j(H) = \lambda_j(H)\boldsymbol{v}_j(H) \qquad (4.3)$$

$$P_j(H) = S_{jj}(H) - \sum_{l=1}^{k} \lambda_l(H) v_{lj}^2(H). \qquad (4.4)$$

Any solution $(\boldsymbol{A}(H), \boldsymbol{P}(H))$ of the above equations yields as value for the objective function of (4.1) the sum of the $(p-k)$ eigenvalues of $\boldsymbol{S}(H) - diag(\boldsymbol{P}(H))$ different from $\lambda_1(H), \ldots, \lambda_k(H)$. At the global minimum this value reduces to the sum of the smallest $(p-k)$ eigenvalues of $\boldsymbol{S}(H) - diag(\boldsymbol{P}(H))$.

Let us consider an elliptically symmetric distribution G on $\mathbb{R}^p$ with parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ and density

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(x) = \frac{g((x-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}))}{\sqrt{det(\boldsymbol{\Sigma})}}$$

where the function g has a strictly negative derivative g'. Assume that the factor model (2.2) holds and the functionals $\boldsymbol{T}$ and $\boldsymbol{S}$ are Fisher consistent, i.e. $\boldsymbol{T}(G) = \boldsymbol{\mu}$ and $\boldsymbol{S}(G) = \boldsymbol{\Sigma}$. Then the eigenvalues $[\lambda_1, \ldots, \lambda_k]$ of $\boldsymbol{A}(G) = \boldsymbol{L}(G)\boldsymbol{L}^t(G)$ are Fisher consistent for the eigenvalues $[\eta_1, \ldots, \eta_k]$ of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^t$, the matrix $\boldsymbol{A}(G)$ is Fisher consistent for $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^t$, and $\boldsymbol{P}(G)$ is Fisher consistent for $\boldsymbol{\Psi}$.

To obtain the influence functions $IF(x, \boldsymbol{P}, G)$ and $IF(x, \boldsymbol{L}\boldsymbol{L}^t, G)$, we will first compute the influence functions $IF(x, \lambda_j, G)$ and $IF(x, \boldsymbol{v}_j, G)$. For the scatter estimators $\boldsymbol{S}$ we are interested in, $IF(x, \boldsymbol{S}, G)$ is known.

Since $(\lambda_1(G), \boldsymbol{v}_1(G)), \ldots, (\lambda_k(G), \boldsymbol{v}_k(G))$ are eigenvalue/eigenvector pairs of the symmetric matrix $\boldsymbol{S}(H) - diag(\boldsymbol{P}(H))$, it is possible to apply lemma 2.1 of Sibson (1979). We use the reformulation of this lemma by Croux and Haesbroeck (2000, lemma 3) yielding

$$IF(x, \lambda_j, G) = \boldsymbol{v}_j^t(G)[IF(x, \boldsymbol{S}, G) - diag(IF(x, \boldsymbol{P}, G))]\boldsymbol{v}_j(G) \tag{4.5}$$

$$\begin{aligned} IF(x, \boldsymbol{v}_j, G) = & \sum_{\substack{l=1 \\ l\neq j}}^{k} \frac{1}{\lambda_l(G) - \lambda_j(G)} \{\boldsymbol{v}_l^t(G)[-IF(x, \boldsymbol{S}, G) + diag(IF(x, \boldsymbol{P}, G))]\boldsymbol{v}_j(G)\}\boldsymbol{v}_l(G) \\ & + \sum_{l=k+1}^{p} \frac{1}{\lambda_j(G) - \lambda_j(G)} \{\boldsymbol{a}_l^t(G)[diag(IF(x, \boldsymbol{P}, G)) - IF(x, \boldsymbol{S}, G)]\boldsymbol{v}_j(G)\}\boldsymbol{a}_l(G) \\ & \sum_{\substack{l=1 \\ l\neq j}}^{k} \frac{1}{\lambda_l(G) - \lambda_j(G)} \{\boldsymbol{v}_l^t(G)[-IF(x, \boldsymbol{S}, G) + diag(IF(x, \boldsymbol{P}, G))]\boldsymbol{v}_j(G)\}\boldsymbol{v}_l(G) \\ & + \sum_{l=k+1}^{p} \frac{-1}{\lambda_j(G)} \{\boldsymbol{a}_l^t(G)[diag(IF(x, \boldsymbol{P}, G)) - IF(x, \boldsymbol{S}, G)]\boldsymbol{v}_j(G)\}\boldsymbol{a}_l(G). \end{aligned} \tag{4.6}$$

The vectors $\boldsymbol{a}_{k+1}(G), \ldots, \boldsymbol{a}_p(G)$ are eigenvectors associated with the $(p-k)$ zero eigenvalues of $\boldsymbol{S}(G) - diag(\boldsymbol{P}(G))$ and form an orthonormal basis of the orthogonal complement of $\boldsymbol{v}_1(G), \ldots, \boldsymbol{v}_k(G)$ in $\mathbb{R}^p$. From equation (4.4) we find the expression of $IF(x, \boldsymbol{P}, G)$:

$$\begin{aligned} IF(x, P_j, G) = & IF(x, S_{jj}, G) - \sum_{l=1}^{k} IF(x, \lambda_l, G) v_{lj}^2(G) \\ & - \sum_{l=1}^{k} 2\lambda_l(G) v_{lj}(G) IF(x, v_{lj}, G). \end{aligned} \tag{4.7}$$

This expression contains the influence functions of λ_l and v_{lj}, so we substitute (4.5) and (4.6) in (4.7). This yields p linear equations with the unknowns $IF(x, P_j, G)$ for $j = 1, \ldots, p$. This system of linear equations can be written as

$$(\boldsymbol{I}_p - \boldsymbol{B})IF(x, \boldsymbol{P}, G) = \boldsymbol{b}(x) \tag{4.8}$$

in which $\boldsymbol{B}$ does not depend on x and $\boldsymbol{b}(x)$ depends on x through $IF(x, \boldsymbol{S}, G)$. Expressions for $\boldsymbol{B}$ and $\boldsymbol{b}(x)$ are derived in the Appendix.

Once we have solved (4.8) for the $IF(x, P_j, G)$ we can easily compute $IF(x, \lambda_j, G)$ and $IF(x, \boldsymbol{v}_j, G)$ from (4.5) and (4.6). By (4.2) this also yields

$$IF(x, \boldsymbol{L}\boldsymbol{L}^t, G) \;=\; IF(x, \boldsymbol{A}, G) \;=\; IF(x, \sum_{j=1}^{k} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^t, G) \tag{4.9}$$

$$\begin{aligned} &= \sum_{j=1}^{k} \{IF(x, \lambda_j, G)\boldsymbol{v}_j(G)\boldsymbol{v}_j^t(G) + \lambda_j(G)IF(x, \boldsymbol{v}_j, G)\boldsymbol{v}_j^t(G) \\ &\quad + \lambda_j(G)\boldsymbol{v}_j(G)IF(x, \boldsymbol{v}_j, G)^t\}. \end{aligned} \tag{4.10}$$

Let us now compare the influence functions of the classical principal factor analysis and the robust principal factor analysis. The difference is due to the $IF(x, \boldsymbol{S}, G)$ of the estimator $\boldsymbol{S}$ being used. The influence function of the classical covariance matrix is

$$IF(x, \boldsymbol{S}^c, G) \;=\; (x - \mu)(x - \mu)^t - \boldsymbol{\Sigma}. \tag{4.11}$$

The influence function of the MCD scatter matrix was derived in (Croux and Haesbroeck 1999) for a distribution G_0 with $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. When working with general $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we use the affine equivariance of $\boldsymbol{S}^r$, yielding

$$IF(x, \boldsymbol{S}^r, G) \;=\; (\boldsymbol{S}^r)^{1/2} IF[(\boldsymbol{S}^r)^{-1/2}(x - \boldsymbol{T}), \boldsymbol{S}^r, G_0](\boldsymbol{S}^r)^{1/2}.$$

The MCD functional $\boldsymbol{S}^r$ depends on the value $0 \leq \alpha \leq 0.5$, where $1 - \alpha \cong h/n$ is the coverage percentage. As in the previous section, we set $\alpha = 0.25$ to obtain a good compromise between efficiency and robustness.

Example 2. Let the trivariate data distribution G be elliptically symmetric with location vector $\boldsymbol{\mu} = \boldsymbol{0}$ and scatter matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Since $p = 3$ we can determine only one factor ($k = 1$). The loadings matrix $\mathbf{\Lambda}$ is $[1, 1, 1]^t$ and the specific variances are given by $\mathbf{\Psi} = [1, 1, 1]$. The influence functions (4.8) and (4.9) can now be computed. Figure 4 shows plots of the classical and robust influence functions. The graphs are made for $x = (x_1, x_2, 0)$ in order to represent them in a three-dimensional plot. (Plots of IF(x_1, x_2, c) for $c \neq 0$ look quite similar.) The influence function $IF(x; P_1^c, G)$ in Figure 4a is unbounded, and shows that an outlying x can have an arbitrarily large effect on $\boldsymbol{P}^c$, confirming the findings of Tanaka and Odaka (1989). On the other hand, the influence function of our robust counterpart in Figure 4b is bounded. Inside the elliptical central region of the x-distribution (corresponding to the MCD) the IF looks like that of the classical PFA in Figure 4a, and outside that region it is constant. Figures 4c and 4d plot the influence function of $(\boldsymbol{L}\boldsymbol{L}^t)_{33}$ for the classical and the robust PFA methods, with the same relation between them. This shows that any outlier x has only a bounded effect on the robust PFA results, no matter how far x is away from G.

In order to obtain smooth influence functions, it suffices to replace the MCD scatter matrix by an S-estimator of multivariate location and scatter (see Rousseeuw and Leroy 1987). These estimators currently need more computation time than the MCD, especially for large n, but their influence function is smooth as can be seen in (Croux and Haesbroeck 1999). We then have to insert the latter influence function into (4.5) – (4.7), yielding smooth versions of the plots in Figure 4.

Until now we considered the IF of PFA based on a covariance matrix. Another possibility is to work with a correlation matrix $\boldsymbol{\rho}$. As in (2.3), this $\boldsymbol{\rho}$ is obtained by the formula $\mathbf{\Sigma}_D^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}_D^{-1/2}$ where $\mathbf{\Sigma}_D$ consists of the diagonal of $\mathbf{\Sigma}$ and zeroes elsewhere. Then the loadings matrix $\mathbf{\Lambda} \in I\!R^{p\times k}$ and the specific variances $\mathbf{\Psi} \in I\!R^p$ satisfy $\boldsymbol{\rho} = \mathbf{\Lambda}\mathbf{\Lambda}^t + diag(\mathbf{\Psi})$. We find analogous equations for $IF(x, \boldsymbol{P}, G)$ and $IF(x, \boldsymbol{L}\boldsymbol{L}^t, G)$, with the only difference that $\boldsymbol{S}(G)$ is replaced by $\boldsymbol{R}(G)$ and therefore $\boldsymbol{v}_j$ and λ_j change. The formula for differentiating a product of three matrices yields

$$\begin{aligned} IF(x, \boldsymbol{R}, G) &= \mathbf{\Sigma}_D^{-1/2} IF(x, \boldsymbol{S}, G)\mathbf{\Sigma}_D^{-1/2} - \frac{1}{2}\mathbf{\Sigma}_D^{-1} IF(x, \boldsymbol{S}_D, G)\boldsymbol{\rho} \\ &\quad -\frac{1}{2}\boldsymbol{\rho}\mathbf{\Sigma}_D^{-1} IF(x, \boldsymbol{S}_D, G). \end{aligned} \tag{4.12}$$

In the bivariate situation, Devlin et al. (1975) gave the influence function of the classical correlation and plotted its contours.

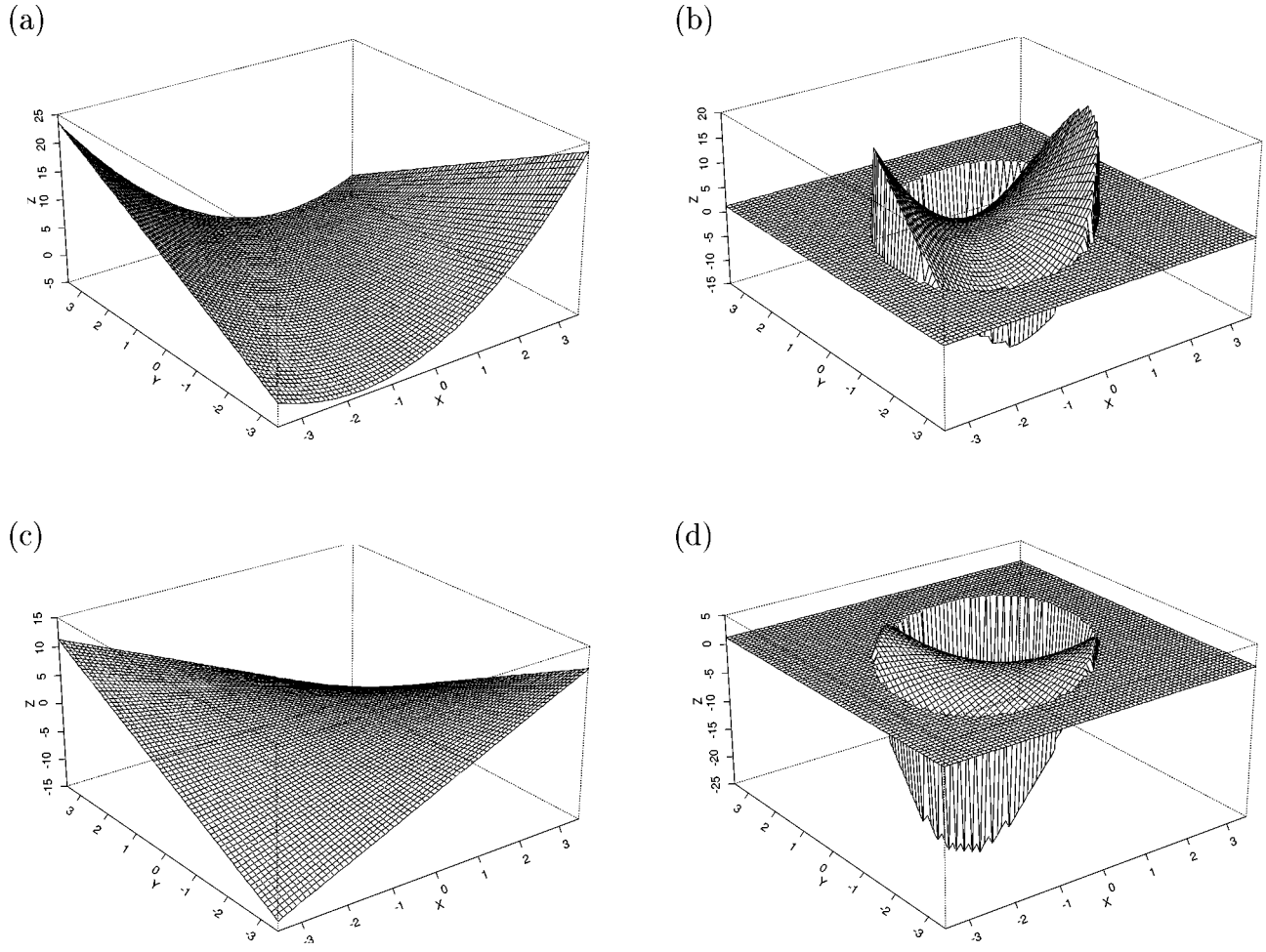

Figure 4: Influence function $IF(x, P_1, G)$ based on (a) the classical covariance matrix and (b) the MCD75 scatter matrix; plot of $IF(x, (LL^t)_{33}, G)$ based on (c) the classical covariance matrix and (d) the MCD75 scatter matrix.

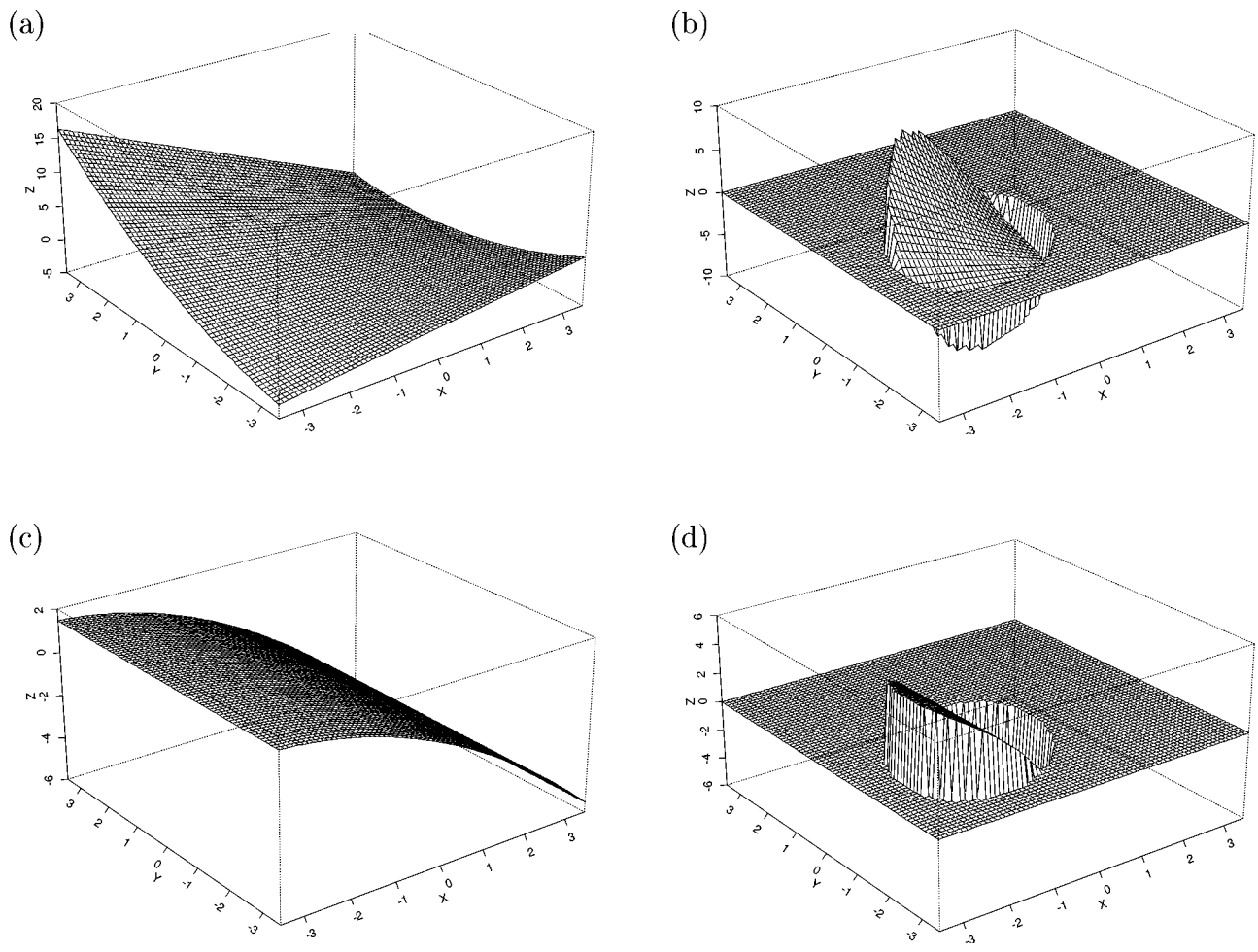

Figure 5: Influence function $IF(x, P_2, G)$ based on (a) the classical correlation matrix and (b) the MCD75 correlation matrix; plot of $IF(x, (LL^t)_{13}, G)$ based on (c) the classical correlation matrix and (d) the MCD75 correlation matrix.

Example 3. We carry out a factor analysis based on the correlation matrix, at the distribution G of the previous example. The population correlation matrix is

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}.$$

The number of factors remains $k = 1$, and now $\boldsymbol{\Lambda} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]^t$ with $\boldsymbol{\Psi} = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]$. Figure 5 shows the influence function of the classical and the robust PFA. The differences between them can be interpreted in roughly the same way as in Figure 4. However, there is an important difference: the constant part in Figures 5b and 5d is zero, whereas that in Figures

4b and 4d is not.

When G_0 is such that $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ the influence function (4.11) of the classical covariance matrix is

$$IF(x, \boldsymbol{S}^c, G_0) \;=\; xx^t - \boldsymbol{I}_p$$

whereas that of the MCD scatter matrix equals

$$IF(x, \boldsymbol{S}^r, G_0) = c\; xx^t I(\|x\| \leq q_\alpha) + w(\|x\|)\boldsymbol{I}_p, \tag{4.13}$$

where w is a certain real-valued function, $q_\alpha = \sqrt{\chi^2_{p,1-\alpha}}$ and c is a constant which depends on α and p, as shown by Croux and Haesbroeck (1999). Therefore the influence functions of $\boldsymbol{S}^c$ and $\boldsymbol{S}^r$ look similar for $\|x\| \leq q_\alpha$ whereas for $\|x\| > q_\alpha$ that of $\boldsymbol{S}^r$ only depends on $\|x\|$.

The influence function of the diagonal elements of the correlation matrix (always ones) is zero. For the off-diagonal elements we only have to consider the first part of the right hand side of expression (4.13). Together with expression (4.12) we obtain

$$IF(x, \boldsymbol{R}^r, G_0) = c\; IF(x, \boldsymbol{R}^c, G_0) I(\|x\| \leq q_\alpha).$$

For general $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the result follows from equivariance:

$$IF(x, \boldsymbol{R}^r, G) \;=\; h(x) IF(x, \boldsymbol{R}^c, G)$$

with $h(x) = c\; I(\|\boldsymbol{\Sigma}^{-1/2}(x - \boldsymbol{\mu})\| \leq q_\alpha)$. From (4.5) to (4.7) it follows that

$$\begin{aligned}
IF(x, \boldsymbol{P}^r, G) &= h(x) IF(x, \boldsymbol{P}^c, G) \\
IF(x, \lambda_j^r, G) &= h(x) IF(x, \lambda_j^c, G) \\
IF(x, \boldsymbol{v}_j^r, G) &= h(x) IF(x, \boldsymbol{v}_j^c, G) \\
IF(x, (\boldsymbol{L}\boldsymbol{L}^t)^r, G) &= h(x) IF(x, (\boldsymbol{L}\boldsymbol{L}^t)^c, G)
\end{aligned}$$

Hence, for factor analysis based on correlations the robust influence functions are 'skipped' versions of the classical influence functions.

4.2 The Empirical Influence Function

Until now 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 can 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 a data point x_i to measure its effect on the principal factor analysis. Our aim is to detect the most influential observations x_i by comparing the EIF(x_i) for $i = 1, \ldots, n$.

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 can 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 four 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 is the simplest, but often masks outliers when there is more than one, because $\boldsymbol{T}_n^c$ and $\boldsymbol{S}_n^c$ break down.

- Masking also occurs with $EIF(x_i; \boldsymbol{P}_n^r; \boldsymbol{T}_n^c, \boldsymbol{S}_n^c)$ for the same reason. We will not consider this possibility further.

- 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. Therefore, we prefer this approach to reveal influential points (i.e. points that would strongly affect the classical PFA). 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 we therefore recommend to compute the $EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)$.

Example 4. Let us illustrate these approaches on the aircraft data set of Example 1. We 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 6 plots $\|EIF(x_i; \boldsymbol{P})\|$ versus the case number i.

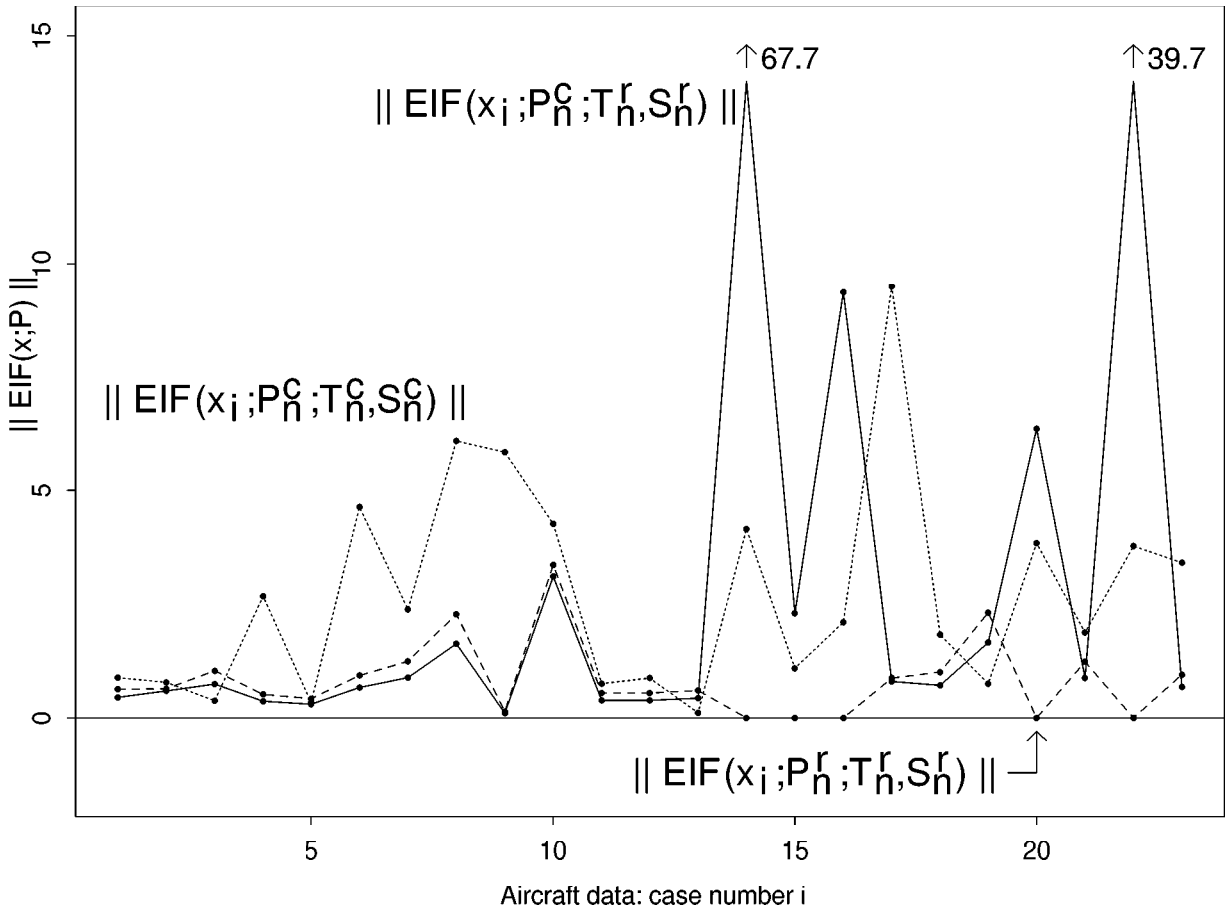

Figure 6: The empirical influence functions $\|EIF(x_i; \boldsymbol{P})\|$ evaluated in 23 aircraft.

We see that the outlying cases 14 and 22 have a 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 (see also Rousseeuw and Van Zomeren 1990). 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.

5 Examples

To illustrate robust factor analysis we consider two real data examples. The vole data set (Airoldi and Hoffmann, 1984) consists of 45 Microtus ochrogaster species. The variables are the age in days (X1), the condylo-incisive length (X2), the length of the incisive foramen

(X3), the alveolar length of the upper molar tooth row (X4) and the interorbital width (X5).

First, we compute the Mahalanobis distances and the robust distances. The robust distances (Rousseeuw and Leroy 1987) are given by

$$RD(x_i) \;=\; d(x_i, \boldsymbol{T}_n^r, \boldsymbol{S}_n^r) \;=\; \sqrt{(x_i - \boldsymbol{T}_n^r)^t (\boldsymbol{S}_n^r)^{-1} (x_i - \boldsymbol{T}_n^r)} \tag{5.1}$$

whereas the Mahalanobis distances $MD(x_i)$ equal $d(x_i, \boldsymbol{T}_n^c, \boldsymbol{S}_n^c)$. As proposed by Rousseeuw and Van Driessen (1999), Figure 7 plots the $RD(x_i)$ versus $MD(x_i)$ with cutoff value $\sqrt{\chi^2_{5,0.975}} \approx 3.58$ on both axes. The robust distances detect eight outliers (cases 3, 4, 8, 9, 23, 39, 40, 41) while the $MD(x_i)$ do not flag any. Let us compute the empirical influence

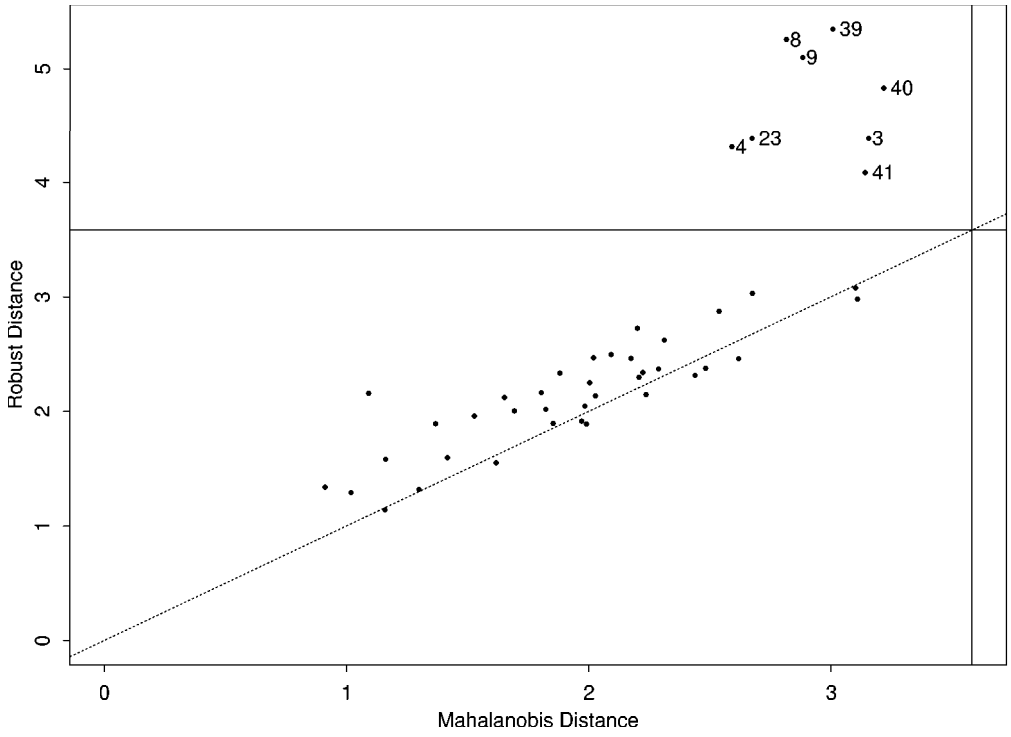

Figure 7: Distance-distance plot of the vole data set.

function $EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)$ for a principal factor analysis with $k = 2$. To see which observations are unusually influential, we need a cutoff value. This value will depend on the data set, because factor analysis is not affine equivariant. (If we transform the data linearly, we cannot simply derive the new loadings and specific variances from the old ones).

To compute the cutoff value we generate data sets $\boldsymbol{X}^{(s)}$ for $s = 1, \ldots, m$ with the same

dimensions, according to the factor analysis model

$$\boldsymbol{X}^{(s)} = \boldsymbol{\Lambda}\boldsymbol{\Phi}^{(s)} + \boldsymbol{\varepsilon}^{(s)}$$

where $\boldsymbol{\Lambda}$ is set equal to the robust estimate $\boldsymbol{L}_n^r$ of the original data, the entries of $\boldsymbol{\Phi}^{(s)}$ are generated from $N(0,1)$, and the entries $\varepsilon_{ij}^{(s)}$ are generated from $N(0,(\boldsymbol{P}_n^r)_j)$. Next, we compute the value $\|EIF(x_i;\boldsymbol{P}_n^c;\boldsymbol{T}_n^r,\boldsymbol{S}_n^r)\|$ for each case x_i in each data set $\boldsymbol{X}^{(s)}$. The cutoff is then obtained as the 95% quantile of all these values. For the vole data we found the cutoff value 23.5. In Figure 8 we see that cases 8, 9, 39, 40, and 41 have an exceptionally high $\|EIF(x_i;\boldsymbol{P}_n^c;\boldsymbol{T}_n^r,\boldsymbol{S}_n^r)\|$, hence these cases are highly influential.

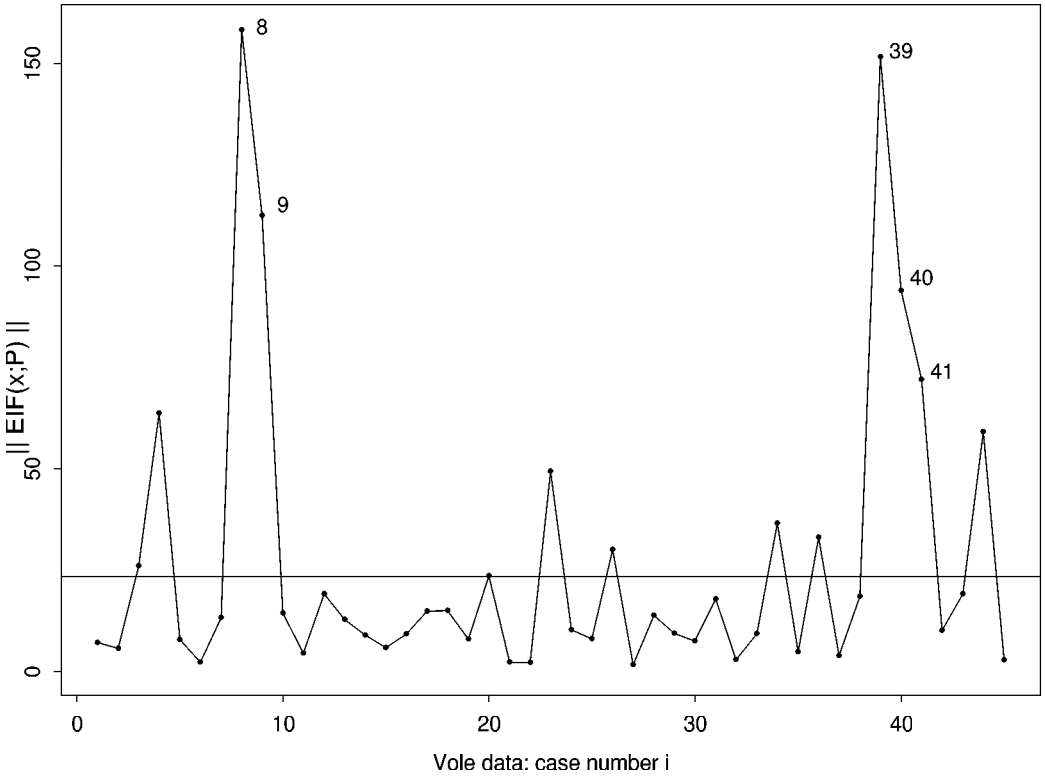

Figure 8: Empirical influence function $\|EIF(x_i;\boldsymbol{P}_n^c;\boldsymbol{T}_n^r,\boldsymbol{S}_n^r))\|$ of the vole data.

Figure 9 shows the biplots of the classical analysis and the robust analysis. As before, the classical factor analysis has the disadvantage that the estimates for $\boldsymbol{\mu}$ and the correlation matrix $\boldsymbol{\rho}$ are affected by the outliers. Therefore the factors and loadings do not give the structure of the correlation matrix of the good objects, since they are also influenced by the outliers. The two biplots are clearly different, due to the differences between $\boldsymbol{R}_n^c$ and $\boldsymbol{R}_n^r$. (For instance, the classical correlation between the variables X3 and X4 is 0.45 and

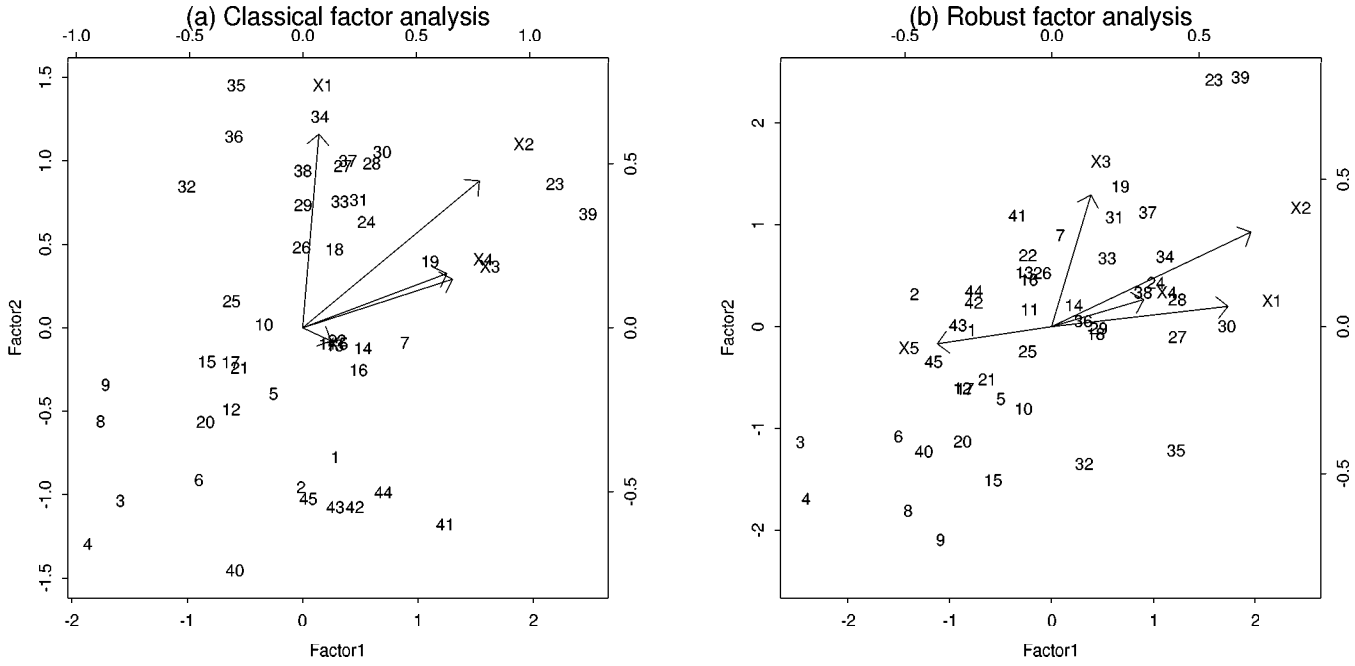

Figure 9: Biplots of (a) classical and (b) robust factor analysis on the vole data.

the robust correlation is 0.12. For the correlation between X2 and X5 we have 0.09 for the classical and -0.35 for the robust correlation.) Also note that cases 36 and 40 have a different position in the two biplots.

Looking at the classical results in Table 2, we see that the variables X2, X3, and X4 load highly on factor 1, and the variables X1 and X2 dominate factor 2. For robust PFA the variables X1, X2, and X5 load highly on factor 1 and the variables X2, and X3 load highly on factor 2. This again illustrates that the robust FA finds a different structure, which in fact corresponds to the data set without the outliers.

Table 2: Loadings of both factor analyses on the vole data.

	Loadings of Classical FA		Loadings of Robust FA	
Variable	Factor 1	Factor 2	Factor 1	Factor 2
X1	0.000	0.750	0.657	0.102
X2	0.791	0.568	0.742	0.477
X3	0.671	0.188	0.147	0.666
X4	0.646	0.210	0.344	0.137
X5	0.126	0.000	-0.426	0.000

The Swiss bank notes data (Flury and Riedwyl 1988) describe 100 forged bank notes of 1000 francs. The variables are the length of the bill (X1), the height of the bill measured on the left (X2), the height of the bill measured on the right (X3), the distance of the inner frame to the lower border (X4), the distance of the inner frame to the upper border (X5) and the length of the diagonal (X6). In the distance-distance plot (Figure 10) the robust distances RD_i detect 19 outliers.

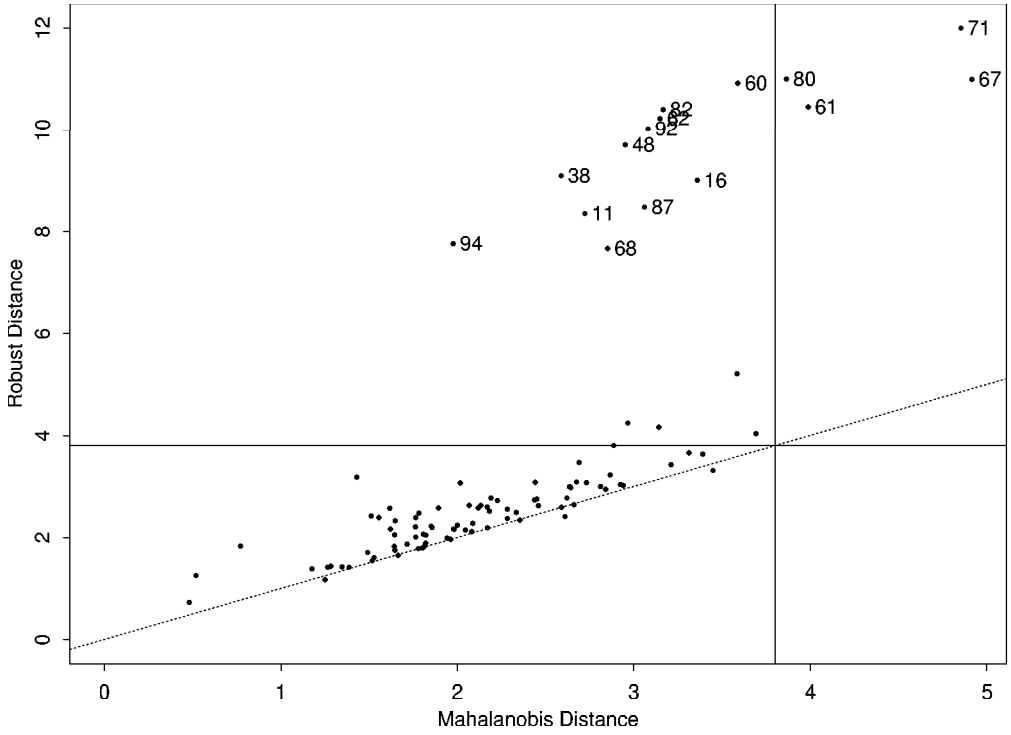

Figure 10: Distance-distance plot of the bank notes data.

For the factor analysis with $k = 2$ the empirical influence function $EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)$ is shown in Figure 11, with the cutoff value 5.99 obtained through simulation. The points with high influence are cases 11, 38, 48, 60, 61, 62, 67, 68, 71, 80, 82, 87, 92 and 94. All of these are also x-outliers, as we can see in Figure 10. However, one of the far x-outliers (case 16) in Figure 10 has only a small influence on the factor analysis (Figure 11). This situation is similar to a bivariate scatterplot, where a point may be far from the data cloud without influencing the regression line. Think of a point lying on the linear trend of the bulk of the data. In regression analysis, this is called a 'good leverage point' (Rousseeuw and Van Zomeren 1990). We could detect such points in factor analysis by

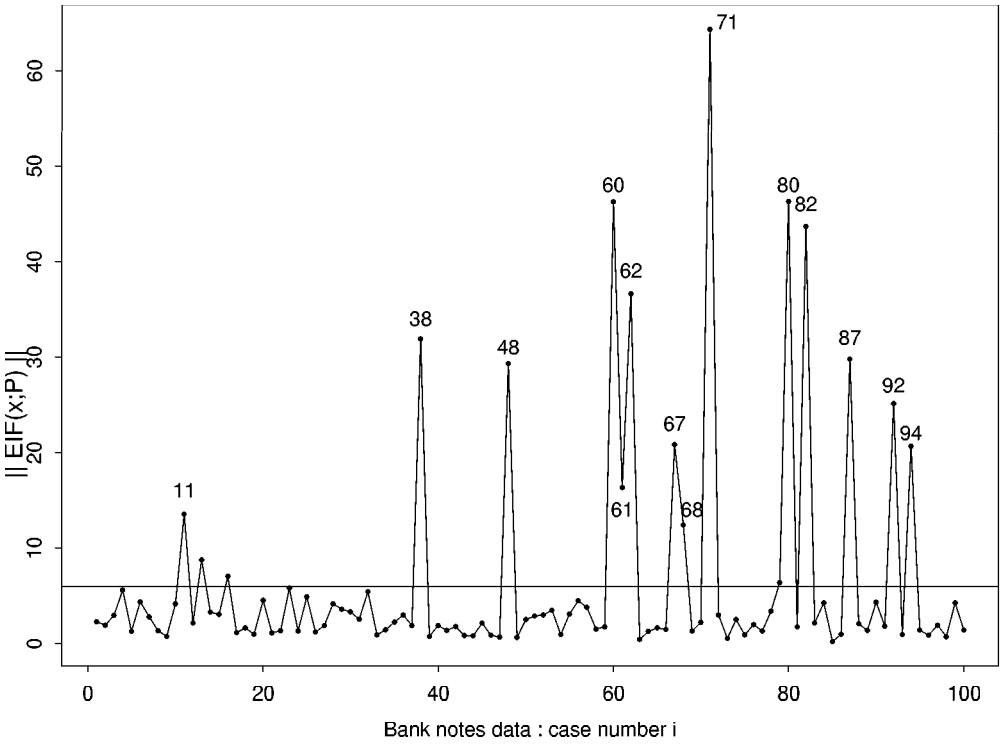

Figure 11: Empirical influence function $\|EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)\|$ of the 100 bank notes.

plotting $\|EIF(x_i; \boldsymbol{P}_n^c; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r))\|$ versus $RD(x_i)$, together with their cutoff values. This would be a useful diagnostic plot.

Let us compare the biplots (Figure 12) and the loadings (Table 3) of the two factor analyses. Variable $X6$ has a different position in the two biplots. This has to do with the fact that the classical correlation between $X1$ and $X6$ is only 0.05, whereas their robust correlation is 0.36. The classical and robust loadings in Table 3 also differ substantially.

6 Discussion

A referee asked to show that our method can also resist outliers in factor space, in the following way. Let us again consider again Table 1. The loading matrix based on classical FA is denoted by $\boldsymbol{\Lambda}_1 \in I\!R^{5\times 2}$, and the one based on the robust FA is denoted as $\boldsymbol{\Lambda}_2$. We now generate 95 data points x_i from the first factor model

$$x_i = \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_i + \boldsymbol{\varepsilon}_i \tag{6.1}$$

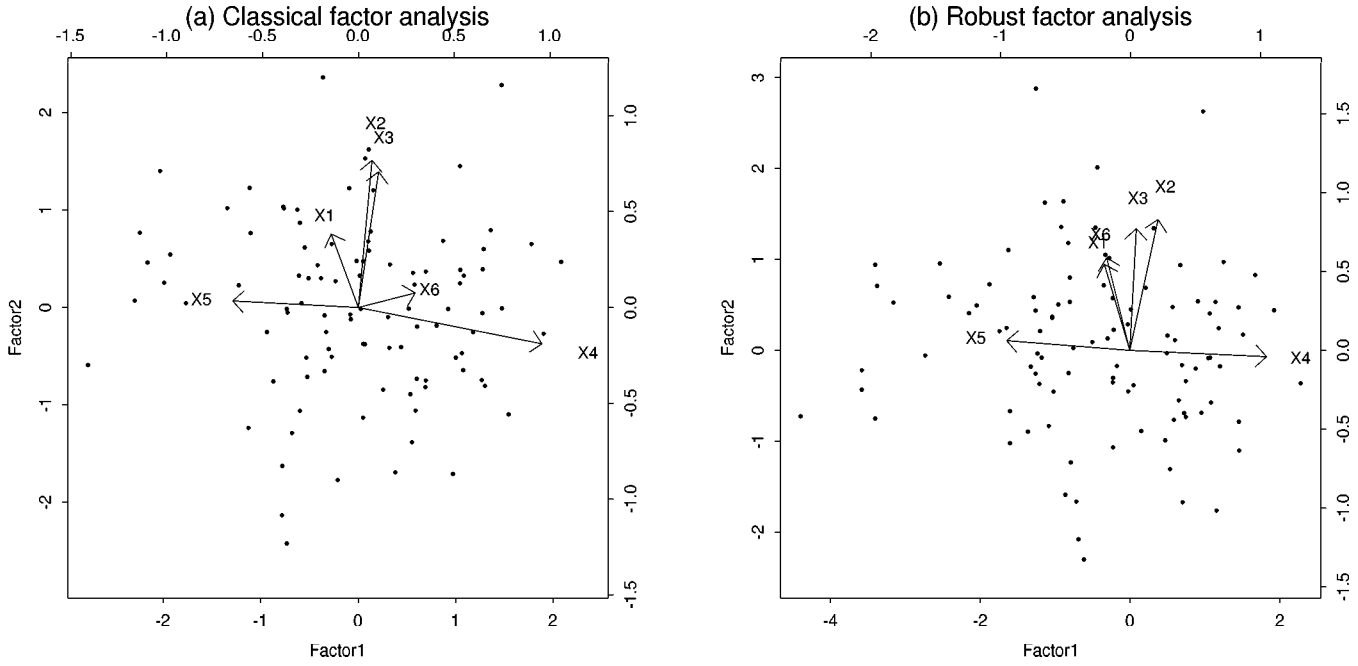

Figure 12: Biplots of the bank notes data: (a) classical, and (b) robust.

with $\boldsymbol{\Phi}_i \sim N_2(0, \boldsymbol{I})$ and $\boldsymbol{\varepsilon}_i \sim N_5(0, \boldsymbol{I})$. We then add 5 additional points to this data set generated from another factor model

$$x_i = \boldsymbol{\Lambda}_2 \boldsymbol{\Phi}_i + \boldsymbol{\varepsilon}_i \tag{6.2}$$

with $\boldsymbol{\Phi}_i$ and $\boldsymbol{\varepsilon}_i$ generated as before. We also checked that the Mahalanobis distances $x_i(\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1^t + \boldsymbol{I})^{-1}x_i^t$ of these 5 additional points were larger than the cutoff value $\chi_5^2(0.975)$ so that these 5 observations deviate from the factor model (6.1).

The empirical influence function $\|EIF(x_i; \boldsymbol{P}_n^r; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)\|$ is plotted in Figure 13. From this plot we can clearly see that the robust method has indeed downweighted these 5 points.

Standard errors for the loading estimates based on the MCD scatter matrix can be computed as follows. Since the MCD is asymptotically normal, see (Butler et al. 1993) and (Croux and Haesbroeck 1999), it follows that under the model the loading matrix $\boldsymbol{L} = [\sqrt{\lambda_1}\boldsymbol{v}_1, \ldots, \sqrt{\lambda_k}\boldsymbol{v}_k]$ which follows the model satisfies

$$\sqrt{n}(\boldsymbol{L}_j - \boldsymbol{\Lambda}_j)^p \longrightarrow N_p(\boldsymbol{0}, ASV(\boldsymbol{L}_j))$$

where $ASV(\boldsymbol{L}_j) = E_G[IF(x, \boldsymbol{L}_j, G)IF(x, \boldsymbol{L}_j, G)^t]$. Using the expressions (4.5) and (4.6) for $IF(x, \lambda_j, G)$ and $IF(x, \boldsymbol{v}_j, G)$ we can obtain the influence function for the vector of loadings $\boldsymbol{L}_j$ as

$$IF(x, \boldsymbol{L}_j, G) = \frac{1}{2\sqrt{\lambda_j}} IF(x, \lambda_j, G) + IF(x, \boldsymbol{v}_j, G)\sqrt{\lambda_j}.$$

Table 3: Loadings of both factor analyses on the bank notes data.

	Loadings of Classical FA		Loadings of Robust FA	
Variable	Factor 1	Factor 2	Factor 1	Factor 2
X1	-0.143	0.403	0.182	0.517
X2	0.000	0.807	-0.202	0.787
X3	0.109	0.744	0.000	0.732
X4	0.974	-0.199	-0.974	0.000
X5	-0.664	0.000	0.879	0.000
X6	0.302	0.000	0.167	0.557

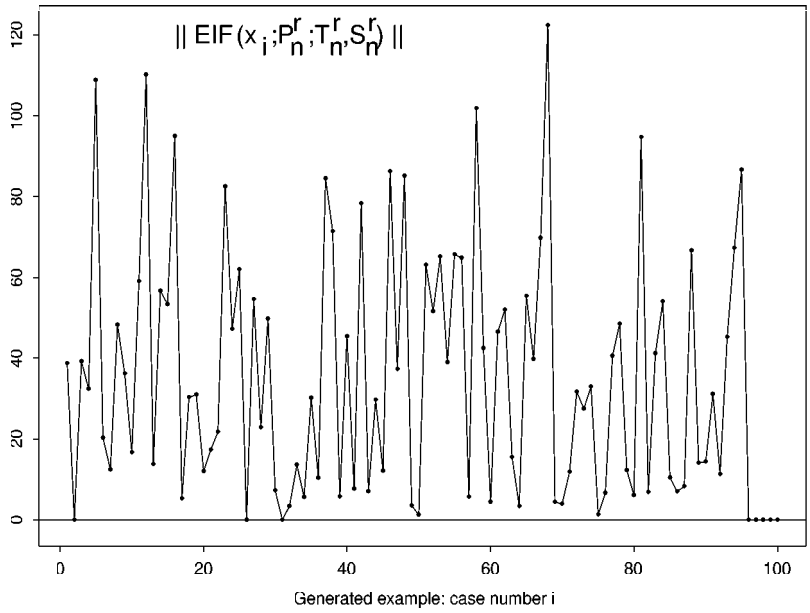

Figure 13: The empirical influence function $\|EIF(x_i; \boldsymbol{P}_n^r; \boldsymbol{T}_n^r, \boldsymbol{S}_n^r)\|$ evaluated in 100 generated points

The covariance matrix of $\boldsymbol{L}_j$ can then be obtained by
$COV(\boldsymbol{L}_j) = \frac{1}{n^2} \sum_{i=1}^{n} IF(x, \boldsymbol{L}_j, \hat{F}_n) IF(x, \boldsymbol{L}_j, \hat{F}_n)^t$ where $\hat{F}_n$ is the empirical distribution. The standard errors can now be obtained as $std(l_ij) = \sqrt{COV(\boldsymbol{L}_j)_{ii}}$. Croux and Dehon (2001) used the same approach to obtain standard errors for robust canonical correlations.

7 Appendix

We will derive the system (4.8) of linear equations. Substituting (4.5) in the right hand side of (4.7) gives

$$IF(x,P_j,G) = IF(x,S_{jj},G) - \sum_{l=1}^{k} v_{lj}^2(G)[\boldsymbol{v}_l^t(G)IF(x,\boldsymbol{S},G)\boldsymbol{v}_l(G)]$$
$$+\sum_{l=1}^{k} v_{lj}^2(G)\boldsymbol{v}_l^t(G)diag[IF(x,\boldsymbol{P},G)]\boldsymbol{v}_l(G) - 2\sum_{l=1}^{k}\lambda_l(G)v_{lj}(G)IF(x,v_{lj},G). \quad (7.1)$$

Since $IF(x,\boldsymbol{S},G)$ is known, (7.1) relates the influence functions $IF(x,\boldsymbol{v}_j,G)$ and $IF(x,P_j,G)$ to each other. Simplifying,

$$IF(x,P_j,G) = IF(x,S_{jj},G) - \sum_{l=1}^{k} v_{lj}^2(G)[\boldsymbol{v}_l^t(G)IF(x,\boldsymbol{S},G)\boldsymbol{v}_l(G)]$$
$$+\sum_{l=1}^{k} v_{lj}^2(G)\{\sum_{s=1}^{p} v_{ls}^2 IF(x,P_s,G)\} - 2\sum_{l=1}^{k}\lambda_l(G)v_{lj}(G)IF(x,v_{lj},G).$$

Let us define the constants

$$c_j^{(1)} = IF(x,S_{jj},G) - \sum_{l=1}^{k} v_{lj}^2(G)[\boldsymbol{v}_l^t(G)IF(x,\boldsymbol{S},G)\boldsymbol{v}_l(G)] \qquad \text{and} \qquad c_{sj}^{(2)} = \sum_{l=1}^{k} v_{lj}^2 v_{ls}^2$$

yielding

$$IF(x,P_j,G) \;=\; c_j^{(1)} + \sum_{s=1}^{p} c_{sj}^{(2)} IF(x,P_s,G) - 2\sum_{l=1}^{k}\lambda_l(G)v_{lj}(G)IF(x,v_{lj},G). \quad (7.2)$$

We now consider equation (4.6) and define the constant vector

$$c_l^{(3)} \;=\; \sum_{\substack{q=1\\ q\neq l}}^{k} \frac{1}{(\lambda_q(G)-\lambda_l(G))}\boldsymbol{v}_q^t(G)[-IF(x,\boldsymbol{S},G)]\boldsymbol{v}_l(G)\boldsymbol{v}_q(G)$$
$$+\sum_{q=k+1}^{p}\frac{-1}{\lambda_l(G)}\boldsymbol{a}_q^t(G)[-IF(x,\boldsymbol{S},G)]\boldsymbol{v}_l(G)\boldsymbol{a}_q(G).$$

This yields

$$IF(x,v_{lj},G) \;=\; c_{lj}^{(3)} + \sum_{\substack{q=1\\ q\neq l}}^{k}\frac{1}{(\lambda_q(G)-\lambda_l(G))}\{\sum_{i=1}^{p} v_{qi}(G)IF(x,P_i,G)v_{li}(G)\}v_{qj}(G)$$
$$+\sum_{q=k+1}^{p}\frac{-1}{\lambda_l(G)}\{\sum_{i=1}^{p} a_{qi}(G)IF(x,P_i,G)v_{li}(G)\}a_{qj}(G).$$

By means of the constant matrix $\boldsymbol{c}^{(4)} \in I\!R^{p\times p\times p}$ given by

$$c_{lji}^{(4)} = \sum_{\substack{q=1\\ q\neq l}}^{k} \frac{1}{(\lambda_q(G)-\lambda_l(G))} v_{qi}(G)v_{li}(G)v_{qj}(G) + \sum_{q=k+1}^{p} \frac{-1}{\lambda_l(G)} a_{qi}(G)v_{li}(G)a_{qj}(G)$$

we obtain the simple formula

$$IF(x, v_{lj}, G) = c_{lj}^{(3)} + \sum_{i=1}^{p} c_{lji}^{(4)} IF(x, P_i, G). \tag{7.3}$$

We can now substitute (7.3) into (7.2), yielding

$$\begin{aligned} IF(x, P_j, G) &= c_j^{(1)} + \sum_{s=1}^{p} c_{sj}^{(2)} IF(x, P_s, G) - 2\sum_{l=1}^{k} \lambda_l(G)v_{lj}(G)c_{lj}^{(3)} \\ &\quad -2\sum_{i=1}^{p}\{\sum_{l=1}^{k} \lambda_l(G)v_{lj}(G)c_{lji}^{(4)}\} IF(x, P_i, G) \end{aligned}$$

Defining the constants

$$b_j(x) = c_j^{(1)} - 2\sum_{l=1}^{k} \lambda_l(G)v_{lj}(G)c_{lj}^{(3)} \qquad \text{and} \qquad c_{ji}^{(5)} = -2\sum_{l=1}^{k} \lambda_l(G)v_{lj}(G)c_{lji}^{(4)}$$

for $i, j = 1, \ldots, p$ we can write

$$IF(x, P_j, G) = b_j(x) + \sum_{s=1}^{p} c_{sj}^{(2)} IF(x, P_s, G) + \sum_{s=1}^{p} c_{js}^{(5)} IF(x, P_s, G).$$

With the notation $B_{js} = c_{sj}^{(2)} + c_{js}^{(5)}$ we find

$$IF(x, P_j, G) = b_j(x) + \sum_{s=1}^{p} B_{js} IF(x, P_s, G)$$

or in matrix notation:

$$(\boldsymbol{I}_p - \boldsymbol{B}) IF(x, \boldsymbol{P}, G) = \boldsymbol{b}(x). \tag{7.4}$$

This system of p linear equations with the unknowns $IF(x, P_s, G)$ for $s = 1, \ldots, p$ can be solved numerically. The matrix $\boldsymbol{B}$ is given by

$$\begin{aligned} B_{js} = \sum_{l=1}^{k} &[v_{lj}^2(G)v_{ls}^2(G) + \\ &\lambda_l(G)v_{lj}(G)(\sum_{\substack{q=1\\ q\neq l}}^{k} \frac{2}{(\lambda_l(G)-\lambda_q(G))} v_{qs}(G)v_{ls}(G)v_{qj}(G) + \sum_{q=k+1}^{p} \frac{2}{\lambda_l(G)} a_{qs}(G)v_{ls}(G)a_{qj}(G))]. \end{aligned}$$

Note that $\boldsymbol{B}$ does not depend on x, whereas $\boldsymbol{b}(x)$ depends on x through $IF(x, \boldsymbol{S}, G)$.

References

Airoldi, J.-P. and Hoffmann, R.S. (1984), "Age Variation in Voles and its Significance for Systematic Studies," *Occasional Papers of the Museum of Natural History*, University of Kansas, Lawrence, Kansas, 111, 1–45.

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

Browne, M.W. and Shapiro, A. (1988), "Robustness of normal theory methods in the analysis of linear latent variable models," *British Journal of Mathematical and Statistical Psychology*, 41, 193–208.

Butler, R.W., Davies, P.L. and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, 21, 1385–1400.

Croux, C. and Dehon, C. (2001), "Analyse canonique basée sur des estimateurs robustes de la matrice de covariance," *La Revue de Statistique Appliquee*, to appear.

Croux, C. and Haesbroeck, G. (2000), "Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies," *Biometrika*, 87, 603–618.

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.

Davies, P.L. (1987), "Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices", *The Annals of Statistics*, 15,1269–1292.

Devlin, S., Gnanadesikan, R. and Kettenring, J. (1975), "Robust estimation and outlier detection with correlation coefficients," *Biometrika*, 62, 531–545.

Flury, B. and Riedwyl, H. (1988), *Multivariate Statistics: A Practical Approach*, Cambridge University Press.

Gabriel, K.R. (1971), "The biplot graphical display of matrices with applications to principal component analysis," *Biometrika*, 58 (3), 453–467.

Gower, J. and Hand, D. (1996), *Biplots*, Chapman and Hall, New York.

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

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust Statistics: The Approach based on Influence Functions*, New York: John Wiley.

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

Jöreskog, K.G. (1963), *Staitstical Estimation in Factor Analysis*, Almqvist and Wiksell, Stockholm.

Kano, Y. (1998), "Improper Solutions in Exploratory Factor Analysis: Causes and Treatments," *Advances in Data Science and Classification*, Rizzi, A., Vichi, M., and Bock, H.-H. (Eds.), Springer-Verlag, Berlin, 375–382.

Kosfeld, R. (1996), "Robust Exploratory Factor Analysis," *Statistical Papers*, 37, 105–122.

Maronna, R.A. (1976), "Robust M-estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51-67.

Mooijaart, A. and Bentler, P.M. (1991), "Robustness of normal theory statistics in structural equation models," *Statistica Neerlandica*, 45, 159–171.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P.J. (1985), "Multivariate Estimation with High Breakdown Point," in *Mathematical Statistics and Applications, Vol. B,* eds. W. Grossmann, G. Pflug, I. Vincze and W. Wertz, Dordrecht: Reidel, 283–297.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, Wiley-Interscience, New york.

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

Rousseeuw, P.J. and Van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–651.

Sibson, R. (1979), "Studies in the Robustness of Multidimensional Scaling: Perturbational Analysis of Classical Scaling," *Journal of the Royal Statistical Society B*, 41,217–229.

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

Ten Berge, J.M.F. and Kiers, H.A.L. (1991), " A Numerical Approach to the Exact and the Approximate Minimum Rank of a Covariance Matrix," *Psychometrika*, 56, 309–315.

Yuan, K-H. and Bentler, P.M. (1998a), "Structural equation modeling with robust covariances," *Sociological Methodology*, 28, 363–396.

Yuan, K-H. and Bentler, P.M. (1998b), "Robust mean and covariance structure analysis," *British Journal of Mathematical and Statistical Psychology*, 51, 63-88.