Fitting Multiplicative Models by Robust Alternating Regressions

C. Croux P. Filzmoser G. Pison P.J. Rousseeuw

Keywords: Biplot, Alternating regression, Exploratory data analysis, Factor analysis, FANOVA, Median polish, Robustness, Two-way table.

Fitting Multiplicative Models by Robust Alternating Regressions

C. Croux P. Filzmoser G. Pison P.J. Rousseeuw

Address for Correspondence:

Prof. Dr. Peter Filzmoser
Institute of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstraße 8-10
A-1040 Vienna, Austria

Tel.: +43 1 58801 10733
Fax: +43 1 58801 10799
Email: P.Filzmoser@tuwien.ac.at

Fitting Multiplicative Models by Robust Alternating Regressions

C. Croux P. Filzmoser G. Pison P.J. Rousseeuw

Affiliations of the Authors:

C. Croux: Department of Applied Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium.

P. Filzmoser: Institute of Statistics and Probability Theory, Vienna University of Technology, Austria.

G. Pison and P.J. Rousseeuw: Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen, Belgium.

Fitting Multiplicative Models by Robust Alternating Regressions

C. Croux * P. Filzmoser † G. Pison‡ P.J. Rousseeuw ‡

Abstract: In this paper a robust approach for fitting multiplicative models is presented. Focus is on the factor analysis model, where we will estimate factor loadings and scores by a robust alternating regression algorithm. The approach is highly robust, and also works well when there are more variables than observations. The technique yields a robust biplot, depicting the interaction structure between individuals and variables. This biplot is not predetermined by outliers, which can be retrieved from the residual plot. Also provided is an accompanying robust R^2-plot to determine the appropriate number of factors. The approach is illustrated by real and artificial examples and compared with factor analysis based on robust covariance matrix estimators. The same estimation technique can fit models with both additive and multiplicative effects (FANOVA models) to two-way tables, thereby extending the median polish technique.

Keywords: Biplot, Alternating regression, Exploratory data analysis, Factor analysis, FANOVA, Alternating regression, Median polish, Robustness, Two-way table.

*Department of Applied Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium.

†Institute of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria

‡Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Antwerpen, Belgium.

1 Introduction

Factor analysis (FA) is a standard multivariate technique that is routinely used in the social and behavioral sciences. The aim of factor analysis is to understand and summarize the correlation structure of the observable variables $X_1, \ldots, X_p$. For this purpose one assumes the existence of $k < p$ 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

$$X_j = \lambda_{j1} F_1 + \lambda_{j2} F_2 + \ldots + \lambda_{jk} F_k + \varepsilon_j \tag{1.1}$$

for each $1 \leq j \leq p$. The error variables $\varepsilon_1, \ldots, \varepsilon_p$ are assumed to be independent of each other and of the factors. The coefficients λ_{jl} are called the *loadings*, and collected into the matrix of loadings Λ. The variances of the error terms are denoted by $\psi_1, \ldots, \psi_p$ and called the *specific variances* or uniquenesses. We will treat (1.1) as a *multiplicative model*, and study the FA model as the basic multiplicative model.

Using the vector notations $\underset{\sim}{X} = (X_1, \ldots, X_p)^\top$, $\underset{\sim}{F} = (F_1, \ldots, F_k)^\top$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^\top$, the usual conditions on factors and error terms can be written as $E(\underset{\sim}{F}) = E(\varepsilon) = 0$, $\mathrm{Cov}(\underset{\sim}{F}) = I_k$, and $\mathrm{Cov}(\varepsilon) = \Psi$, where $\Psi = \mathrm{diag}\,(\psi_1, \ldots, \psi_p)$ is a diagonal matrix, containing the specific variances on its diagonal. Furthermore, ε and $\underset{\sim}{F}$ are assumed to be independent.

In factor analysis, one needs to estimate the matrix Λ (which is only specified up to an orthogonal transformation) and Ψ. Classical FA methods are very vulnerable to outliers (Tanaka and Odaka 1989a,b), hence more robust methods need to be constructed.

Nearly all FA procedures are based on a decomposition of the covariance matrix Σ of $\underset{\sim}{X}$. Indeed, from $\underset{\sim}{X} = \Lambda \underset{\sim}{F} + \varepsilon$ it follows that

$$\Sigma = \Lambda \Lambda^\top + \Psi. \tag{1.2}$$

In classical factor analysis, the matrix Σ is estimated by the sample covariance matrix $\hat{\Sigma}$. (When $X_1, \ldots, X_p$ are standardized versions of the original variables, Σ becomes the correlation matrix.) Next, one tries to decompose $\hat{\Sigma}$ as in (1.2) to obtain estimates of Λ and Ψ. Typically $\hat{\Sigma}$ cannot be decomposed exactly as in (1.2), so we must resort to an approximate decomposition. 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 (see, e.g., Basilevsky 1994). However, it is well known that outliers can heavily influence the classical estimate of Σ and hence also the parameter estimates. It is therefore natural to insert a robust scatter matrix estimator instead of the sample covariance matrix. This approach was taken by Kosfeld (1996) who inserted a multivariate

M-estimator, and by Filzmoser (1999) who used the Minimum Volume Ellipsoid (MVE) estimator (Rousseeuw 1985) in a geostatistical problem. Since the MVE estimator has a non-normal convergence, Pison et al (2002) instead used the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985). The MCD looks for the subset of h observations out of n having the smallest determinant of its sample covariance matrix. Typically, $h \approx 3n/4$. The MCD estimator for Σ is then a multiple of that covariance matrix. Pison et al (2002) showed that a robust PFA method is preferable to a robust ML approach and that PFA based on MCD results in a factor analysis method with bounded influence function. This kind of approach is conceptually simple and fast, but it is limited to data sets with fewer variables than objects (i.e. $p < n$) which is not always the case. Moreover, it will turn out that the method we will introduce remains an interesting alternative to the approach based on robust covariance matrix estimators when $n \geq p$.

In this paper an approach is proposed which estimates the unknown parameters directly, without passing via an estimate of the covariance matrix. For this we will modify the technique of alternating regression of Wold (1966), also called criss-cross regression by Gabriel and Zamir (1979). The sample version of model (1.1) is given by

$$x_{ij} = \sum_{l=1}^{k} \lambda_{jl} f_{il} + \varepsilon_{ij} \tag{1.3}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let us for a moment consider the *factor scores* f_{il} as fixed or known and suppose that preliminary estimates for them are known. The loadings λ_{jl} can then be estimated by linear regressions of the x_{ij} on the scores. On the other hand, if preliminary estimates of the loadings are available, we can estimate the scores f_{il} by linear regressions of the x_{ij} on the loadings. Our approach will combine these two viewpoints. Moreover, estimates $\hat{\psi}_j$ for ψ_j can easily be obtained from the residuals. In view of possible outliers, all estimations will be done robustly. We propose to use a weighted L^1 regression estimator, which is robust in this setting and can be fastly computed.

The approach we will pursue will be called RAR, from Robust Alternating Regression. It treats the rows and columns of the data matrix in the same way, which we will see is useful for dealing with missing values and outliers. Section 2 defines the RAR estimator and Section 3 describes the algorithm in more detail. Experiments on real and simulated data show that this method works well, converges quickly and is highly robust. A documented S-plus function for RAR is freely available at *http://www.statistik.tuwien.ac.at/public/filz/*. An accompanying robust R^2-plot is presented in Section 4. This R^2-plot helps to select the number of factors in (1.3). Section 5 presents a real and an artificial data example. A robust

biplot is obtained by taking $k = 2$ factors and simultaneously plotting the individuals by $(\hat{f}_{i1}, \hat{f}_{i2})$ and the variables by $(\hat{\lambda}_{j1}, \hat{\lambda}_{j2})$. The robust biplot shows the main features of the data set and is not affected much by outliers in the data. Such outliers can be detected from the robust residuals.

Section 6 describes a simulation comparing the proposed method with classical PFA and other competitors. Section 7 extends the RAR procedure to fit another multiplicative model, the Factor Analysis of Variance (FANOVA) model introduced by Gollob (1968). A FANOVA model combines aspects of ANOVA and factor analysis. In the FANOVA setting, the RAR estimator can be seen as an extension of the well-known median polish technique. Conclusions are formulated in Section 8.

2 The RAR Estimator

As usual, the $n \times p$ data matrix X contains the individuals (cases, objects) in the rows and the observed variables (characteristics) in the columns. The variables are already standardized to have zero location and unit spread. A factor score is denoted as f_{il}. The ith score vector is given by $f_i = (f_{i1}, \ldots, f_{ik})^\top$, while the jth loading vector is $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jk})^\top$. Both the loading vectors and the score vectors are unknown. Denote by $\theta = (f_1^\top, \ldots, f_n^\top, \lambda_1^\top, \ldots, \lambda_p^\top)$ the vector of all scores and loadings, and let

$$\hat{x}_{ij}(\theta) = \sum_{l=1}^{k} f_{il}\lambda_{lj} = f_i^\top \lambda_j = \lambda_j^\top f_i$$

be the fitted value of x_{ij} according to the model (1.3). By choosing θ such that the fitted and the actual values of the data matrix are close together, we define estimates $\hat{f}_i$ for the score vectors and $\hat{\lambda}_j$ for the loading vectors. The fitted data matrix $\hat{X}$ can then be decomposed as

$$\hat{X} = \hat{F}\hat{\Lambda}^\top \tag{2.1}$$

where the rows of $\hat{F}$ are the estimated scores and the rows of $\hat{\Lambda}$ are the estimated loadings. Observe that the rank of $\hat{X}$ is at most $k < p$, while the rank of X is typically p.

The least squares (LS) approach is to minimize the sum of squared residuals:

$$\hat{\theta}_{LS} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij}(\theta))^2. \tag{2.2}$$

The resulting $\hat{X}$ can be seen as the "best" (in the least squares sense) approximation of the data matrix X by a rank k matrix. The Eckart-Young theorem (Gower and Hand 1996,

p. 241) says that this best fit can be obtained by performing a singular value decomposition $X = UDV^\top$ of the data matrix. By replacing all singular values in D by zero except for the k largest ones, one obtains D_k and finally $\hat{X} = UD_kV^\top$. By taking $\hat{F} = \sqrt{n}U$ and $\hat{\Lambda} = VD_k/\sqrt{n}$ we obtain the so-called *Principal Component solution* to the FA problem (cfr. Johnson and Wichern 1998, p. 524). Moreover, the sample covariance matrix of the estimated score vectors equals $\hat{F}^\top\hat{F}/n = I_k$ which is consistent with the assumption $Cov(\underset{\sim}{F}) = I_k$. Note that we are interested in estimating the factor model (1.3), and that we will not be deriving robust principal components (as has been done in Croux and Haesbroeck 2000, for example). In principal components one constructs linear combinations of observed variables, while in a factor model the observed variables are generated by unobserved factors. In other words, in a factor model the observed variables are at the left hand side of the equation, while in a principal components model they are at the right hand side.

It is important to note that the estimates $\hat{F}$ and $\hat{\Lambda}$ in (2.1) are only specified up to a linear transformation. Since $\hat{X} = (\hat{F}T^\top)(\hat{\Lambda}T^{-1})^\top$ for any non singular k by k matrix T, it follows that $\hat{F}T^\top$ and $\hat{\Lambda}T^{-1}$ attain the same value for the objective (2.2). However, the fitted values $\hat{X}$ are uniquely defined. Moreover, if we add the restriction that the estimated covariance matrix of the score vectors needs to be the identity matrix, then the estimates $\hat{F}$ and $\hat{\Lambda}$ in (2.1) are specified up to an orthogonal transformation, making the matrix $\hat{\Lambda}\hat{\Lambda}^\top$ uniquely defined.

Since the LS criterion gives too much weight to large residuals, a first idea is to use the L^1 criterion (or Least Absolute Deviations criterion) instead, which is known to give a very robust additive fit to two-way tables (Terbeck and Davies 1998). This yields the estimator

$$\hat{\theta}_{L1} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{p} |x_{ij} - \hat{x}_{ij}(\theta)|. \tag{2.3}$$

For the optimal $\hat{F}$ and $\hat{\Lambda}$, it must hold that $\hat{f}_i$ minimizes $\sum_{j=1}^{p} |x_{ij} - f_i^\top\hat{\lambda}_j|$ and $\hat{\lambda}_j$ minimizes $\sum_{i=1}^{n} |x_{ij} - \hat{f}_i^\top\lambda_j|$. Therefore, instead of minimizing both sums in (2.3) at the same time, one fixes an index j and scores f_i and selects the λ_j to minimize

$$\sum_{i=1}^{n} |x_{ij} - f_i^\top\lambda_j|. \tag{2.4}$$

The above problem is now linear instead of bilinear and can easily be solved with a Least Absolute Deviations regression algorithm. One sees immediately that minimizing (2.4) consecutively for $j = 1, \ldots, p$ corresponds to minimizing (2.3) for fixed scores. Analogously, for fixed loadings λ_j, finding the f_i minimizing

$$\sum_{j=1}^{p} |x_{ij} - f_i^\top\lambda_j| \tag{2.5}$$

(for each $i = 1, \ldots, n$ in turn) corresponds to minimizing (2.3) when the loadings are given. Alternating (2.4) and (2.5) leads to an iterative scheme of alternating regressions. Note that the value of the criterion in (2.3) decreases at each step.

Similar algorithms, but based on alternating classical least squares regressions, are popular in chemometrics (Martens and Naes 1989) and the behavioral sciences (Gifi 1990). See also (de Falguerolles and Francis 1992, Gabriel 1998) for generalized bilinear models.

Unfortunately, L^1 regression is sensitive to leverage points. If outlying score or loading vectors are present, the L^1 regressions can be heavily influenced by them. By downweighting these leverage points we obtain a weighted L^1 regression, resulting in the estimator

$$\hat{\theta}_{RAR} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{p} w_i(\theta) v_j(\theta) |x_{ij} - \hat{x}_{ij}(\theta)|. \tag{2.6}$$

One single objective function estimates $\hat{F}$ and $\hat{\Lambda}$ simultaneously from the rows and columns of X. The result of (2.6) is named the RAR estimator, since we will use Robust Alternating Regressions to compute it. The estimator will not be misled by outlying observations.

The row weights in (2.6) are defined by

$$w_i(\theta) = \min(1, \chi^2_{k,0.95}/\mathrm{RD}_i^2) \quad \text{for } i = 1, \ldots, n \tag{2.7}$$

where $\chi^2_{k,0.95}$ is the upper 5% critical value of a chi-squared distribution with k degrees of freedom, and

$$\mathrm{RD}_i = \sqrt{(f_i - T(F))^\top C(F)^{-1} (f_i - T(F))} \quad \text{for } i = 1, \ldots, n$$

are robust distances (Rousseeuw and van Zomeren 1990) computed from the collection of score vectors $F = \{f_i | 1 \leq i \leq n\}$ in k-dimensional space. Such weights were used by Simpson et al (1992) and yielded stable results. The robust multivariate location and scatter estimators T and C are taken as the location and scatter part of the MVE estimator (Rousseeuw 1985). The MVE estimator was chosen here since it performs well as an outlier identifier (see Becker and Gather 2001). Analogously, the set of column weights v_j is defined using the loading vectors. Note that, since the true loadings and scores are unobserved, w_i and v_j depend on the unknown parameter vector θ.

From the robust residuals $\hat{\varepsilon}_{ij} = x_{ij} - \hat{x}_{ij} = x_{ij} - \hat{f}_i^\top \hat{\lambda}_j$, we can estimate the specific variances using

$$\hat{\psi}_j = (\mathrm{MAD}_j(\hat{\varepsilon}_{ij}))^2. \tag{2.8}$$

Here, MAD is made consistent at univariate normal distributions by multiplication with 1.4826, so that it will estimate the same quantity as the nonrobust standard deviation.

Note that the estimates $\hat{\psi}_j$ are positive by construction, so there are never problems with negatively estimated specific variances.

The specific variances do not come in explicitly in the definition of the estimators for the loadings, and are estimated at the very end. This is in contrast with most other factor analysis procedures. Our experiments indicated that, for reasons of robustness, it is better not to make an extra heteroscedasticity weighting in the criss-cross regression scheme. If the regression estimators are consistent in presence of heteroscedasticity, the procedure will maintain its validity even for Ψ not proportional to the identity matrix. Recall that LS estimators and also many robust regression estimators remain consistent under not too heavy forms of heteroscedasticity (see e.g. El Bantli and Hallin (1999) for L^1-estimators and Hallin and Mizera (2001) for M-estimators).

Remark: It was pointed out by the referees that there is a consistency problem for the RAR estimator (as well as for $\hat{\theta}_{L1}$ and $\hat{\theta}_{LS}$). If the scores f_i would be exactly known, then the consistency of the L^1 regression estimator implies consistency for $\hat{\lambda}_j$, since

$$\hat{\lambda}_j = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} w_i |x_{ij} - f_i^\top \beta|.$$

It would then also follow that $\hat{\psi}_j \to \psi_j$ for $n \to \infty$, as long as we use a consistent scale estimator applied on the residuals $x_{ij} - f_i^\top \hat{\lambda}_j$, as we did in (2.8). The problem is of course that the scores f_i are not known but estimated. Only for $p \to \infty$ the estimated scores approach the true ones. In practice the dimension p is finite, and we encounter a "finite dimension bias". Of course, since the sample size n is also finite, we have as well the more familiar "finite sample bias". A more formal study of the asymptotics of the RAR estimator (for both n and p tending to infinity) is beyond the scope of this paper, and may even be infeasible.

3 The RAR Algorithm

The RAR estimator defined in (2.6) can be approximated by an alternating algorithm, as outlined below.

- **Step 0:** To obtain invariance with respect to a change of measurement units, the data are first scaled in a robust way:

$$x_{ij} \leftarrow \frac{x_{ij} - \mathrm{med}_i(x_{ij})}{\mathrm{MAD}_i(x_{ij})}, \tag{3.1}$$

where MAD stands for the Median Absolute Deviation. Note that orthogonal or affine equivariance properties are not necessary in a factor model. This initial standardization corresponds with a correlation matrix based FA. Standardizations using other estimators of location and scale could be envisaged, but we prefer to stick to the traditional choice (3.1).

- **Step 1: starting values.** First, a robust principal component analysis (PCA) procedure is performed. The resulting scores are then taken as starting values $\hat{f}_i^{(0)}$ for the factor scores. We use the projection pursuit (PP) based estimator of Li and Chen (1985), implemented as in Croux and Ruiz-Gazen (1996). This PP-based method is fast to compute, can deal with $p > n$, and is highly robust. Moreover, this approach allows one to compute just the first k principal components (the only ones that are needed here), which reduces the computation time even further. Using classical PCA in this first stage would slow down the convergence considerably, and could lead to a nonrobust FA when there are many outliers. Alternatively, one could take several random starting values, which could help to check for a local versus global optimum. But the latter approach will increase computation time significantly. In any case, experiments have shown that the choice of the starting values is not too crucial for finding a good approximation.

- **Step 2: the iteration process.** Now suppose that the iteration process has reached step t ($t \geq 1$) of the algorithm, and the $\hat{f}_i^{(t-1)}$ are available.

  * First compute weights $w_i^{(t)}$ as defined in (2.7), which downweight outliers in the set of estimated score vectors $\{\hat{f}_i^{(t-1)} | 1 \leq i \leq n\}$ in $\mathbb{R}^k$. Then compute

$$\hat{\lambda}_j^{(t)} = \underset{\lambda \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{n} w_i^{(t)} |x_{ij} - \lambda^\top \hat{f}_i^{(t-1)}| \tag{3.2}$$

  for $j = 1, \ldots, p$. In this part of the procedure, one needs to perform an L^1 fit p times (and this will be the case at every iteration step). Note that the loadings are estimated one at a time, which turned out to be more convenient for the implementation of the algorithm. Fortunately, very efficient algorithms for L^1 regression exist (Bloomfield and Steiger 1983), so this takes little time. Note that the weights $w_i^{(t)}$ only need to be computed once every iteration step. They require computation of a robust scatter estimator in the factor space, which is usually of a low dimension k.

* We analogously compute weights $v_j^{(t)}$ which downweight outliers in the set of estimated loading vectors $\{\hat{\lambda}_j^{(t)} | 1 \leq j \leq p\}$ in $\mathbb{R}^k$. Then compute

$$\hat{f}_i^{(t)} = \underset{f \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{j=1}^{p} v_j^{(t)} |x_{ij} - f^\top \hat{\lambda}_j^{(t)}| \tag{3.3}$$

for $i = 1, \ldots, n$.

* The values of the objective function (2.6) computed for the estimates obtained in step $t-1$ and step t are compared. If there is no essential difference in the objective function, the iterative process is stopped and we set $\hat{f}_i = \hat{f}_i^{(t)}$ for $1 \leq i \leq n$ and $\hat{\lambda}_j = \hat{\lambda}_j^{(t)}$ for $1 \leq j \leq p$. If not, Step 2 is repeated.

- **Step 3: orthogonalization.** This last step is optional and will not alter the fitted values $\hat{X} = \hat{F}\hat{\Lambda}^\top$. We compute a robust estimator $\hat{\Sigma}_f$ of the covariance matrix of the estimated scores $\{\hat{f}_i | 1 \leq i \leq n\}$. Since the scores only have dimension k, where k is small, the matrix $\hat{\Sigma}_f$ can be computed quickly. We compute $\hat{\Sigma}_f$ by the reweighted MCD estimator with 25% breakdown value, using the FAST-MCD algorithm of Rousseeuw and van Driessen (1999). The breakdown value 25% for the MCD has been chosen since this combines robustness with efficiency (see e.g. Croux and Haesbroeck 1999). Afterwards we set

$$\hat{F} \leftarrow \hat{F}\hat{\Sigma}_f^{-1/2} \quad \text{and} \quad \hat{\Lambda} \leftarrow \hat{\Lambda}\hat{\Sigma}_f^{1/2}.$$

  The effect of the above transformation is that the robust covariance matrix of the estimated scores is now an identity matrix, which mimics the model condition $Cov(\underset{\sim}{F}) = I_k$. Another effect is that the biplot representation of the n cases (see Step 4) will show no correlation structure, as is common practice in the biplot literature (Gower and Hand 1996).

- **Step 4: Residuals, uniquenesses, biplot.** The residuals are obtained as $\hat{\varepsilon}_{ij} = x_{ij} - \hat{x}_{ij} = x_{ij} - \hat{f}_i^\top \hat{\lambda}_j$, and can be plotted versus (i, j) in the horizontal plane. This residual plot is very useful for detecting outliers. From the residuals the uniquenesses can be estimated as in (2.8). In the common case $k = 2$ one can represent the individuals by $(\hat{f}_{i1}, \hat{f}_{i2})$ and the variables by $(\hat{\lambda}_{j1}, \hat{\lambda}_{j2})$ in the same 2D plot, called the biplot. Section 5 shows examples of the robust residual plot and the robust biplot.

An S-plus function for the RAR estimator is freely available at *http://www.statistik.tuwien.ac.at/public/filz/*. It also allows to perform alternating regression using other regression estimators, like M-estimators or the highly robust Least Trimmed Squares (LTS) and Least

Median of Squares (LMS) estimators. It is even possible to execute the algorithm with the nonrobust Least Squares regression estimator, yielding the same result as the classical approach of Gabriel (1978) based on the singular value decomposition. Alternating regression using the LMS algorithm was already considered by Ukkelberg and Borgen (1993). However, using the LMS yields a very time consuming algorithm. In our experience, the RAR estimator gave the most satisfying factor analysis method with respect to computation time, robustness, and stable convergence of the algorithm. Although no proof of convergence exists, many simulations and examples have shown its good numerical and statistical performance. It could be mentioned that even the classical FA procedures may have convergence problems.

When the data contain no severe outliers, the unweighted L^1 estimator is a valuable alternative. It is easy to see that for the L^1-based method the objective function (2.3) decreases in each step of the algorithm. The L^1 procedure therefore always converges, although it might happen that the result is not the global minimum. In practice, we found that it always came very close. One could of course use the resulting estimates as starting values for a general purpose optimization procedure for minimizing (2.3). Since the starting value is likely to be very close to the solution, we have a good chance of attaining the global minimum of (2.3).

The RAR procedure required the choice of several auxiliary robust estimators and a weighting function. Most of these choices are standard, and simulations for other robust choices led to essentially identical results.

Remark: It is important to note that it is nowhere required that the number of observations should exceed the number of variables. There is however a restriction on the number of factors k. The computation of the MVE or MCD, required for computing the weights in the weighted L^1 procedure, requires that

$$k < \frac{\min(n, p)}{2}. \tag{3.4}$$

Since dimension reduction is one of the major aims of factor analysis, (3.4) is not a real restriction. (A nice feature of the unweighted L^1 procedure is that it can be computed for k up to the rank of X, which equals $\min(n, p)$.) The robust R^2-plot, which will be presented in the next section, can be used to select an appropriate value for k.

4 A Robust R^2-plot

After having fitted the factor model (1.3) with the weighted L^1 approach, a natural measure of the variability explained by the k factors is

$$R^2_{RAR}(k) = 1 - \left(\frac{\sum_{i=1}^n \sum_{j=1}^p w_i v_j |x_{ij} - \hat{x}_{ij}|}{\sum_{i=1}^n \sum_{j=1}^p w_i v_j |x_{ij}|} \right)^2 . \tag{4.1}$$

The weights are those of (2.7), with the final estimated scores and loading vectors. The definition of the measure $R^2_{RAR}(k)$ resembles the definition of the R^2 measure in classical regression, and compares the dispersion of the residuals in the full model with the dispersion of the residuals in the baseline model without factors. The latter residuals are the observations x_{ij} themselves (recall that the x_{ij} were standardized). Surely, by definition of $\hat{\theta}_{RAR}$, R^2_{RAR} is a number between 0 and 1.

For the L^1-based approach, $R^2_{L1}(k)$ is defined as in (4.1) but with all weights equal to 1. The analogous measure for the LS fit (2.2) is

$$\begin{aligned} R^2_{LS}(k) &= 1 - \frac{\sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \hat{x}_{ij})^2}{\sum_{i=1}^n \sum_{j=1}^p x_{ij}^2} \\ &= 1 - \frac{\text{trace}\left((X - \hat{X})(X - \hat{X})^\top\right)}{\text{trace}(XX^\top)}. \end{aligned}$$

Using the singular value decomposition we find $X = UDV^\top$ and $\hat{X} = UD_kV^\top$ with $D = \text{diag}(\sigma_1, \ldots, \sigma_p)$ where $\sigma_1 \geq \ldots \geq \sigma_p$ and $D_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. Note that the σ_l^2/n are the eigenvalues of the sample correlation matrix. This implies that for the LS fit

$$R^2_{LS}(k) = 1 - \frac{\text{trace}\left((D - D_k)^2\right)}{\text{trace}(D^2)} = \frac{\sum_{l=1}^k \sigma_l^2}{\sum_{l=1}^p \sigma_l^2},$$

which corresponds to the percentage of the total variance explained by the first k factors.

A plot of $R^2_{RAR}(k)$ for a range of values of k will be called a robust R^2-plot. An appropriate value for k can be selected on the basis of this plot, in a similar way as the selection of the number of factors in principal components analysis. Alternatively, one could plot the change in $R^2_{RAR}(k)$ when adding the k th factor to the model. This would resemble the *scree plot* of principal component analysis, and therefore we will call it a robust scree R^2-plot.

5 Examples

In this section we apply RAR to real and artificial data. The datasets can be downloaded from the beforementioned website. We also discuss the robust residual plot, biplot, and R^2-plot.

5.1 European Population Data

Variables related to the health and fertility of a population were measured for 14 European countries and two large groups of countries, the Soviet Union (SU) and the European Community (EU), in their 1986 configuration. The data set is reported in Table 1 and has $n = 16$ and $p = 9$. The variables are average population growth from 1986-2000 (*pop_growth*), percentage of women of the age able to give birth (*give_birth*), proportion of women of all ages per 100 men (*women%*), life expectancy of women (*lifeexp_f*) and men (*lifeexp_m*), infant mortality rate (*inf_mort*), number of inhabitants per physician (*inhab/doc*), daily calorie consumption per head (*calorie*), and proportion of babies with underweight at birth in % (*baby_underw*). The data originate from the European statistical agency EUROSTAT.

Table 1 is inserted about here.

In Figure 1 the R^2-plot and the scree R^2-plot are given for the LS, L^1, and RAR method. The plots indicate that for each method a two-factor model is appropriate, since not much additional variation is explained by using more factors. Note that the first factor for the LS method contributes much more than the second one, while for the robust methods this difference has been smoothed out.

Figure 1 is inserted about here.

After having estimated the two-factor model, we look at the 3D-plots of the residuals $x_{ij} - \hat{x}_{ij}$ versus the pair (i, j) in the horizontal plane. Figure 2 shows these residual plots for classical principal factor analysis (PFA) and RAR. In the classical plot the residuals are very small, and no outliers are visible. Aside from the outliers, the residual plot for the RAR method looks very smooth, but this is only a scale effect. On the other hand, the RAR plot has large residuals for cells (2,2), (2,7) and (14,2), so this method detects AL (Albania) and TR (Turkey) as outliers. In general, a robust approach yields a good fit to the majority of the data. This can be illustrated by computing the sum of squared residuals $\sum_i \sum_j (x_{ij} - \hat{x}_{ij})^2$ with the index i running over all rows except AL and TR. This yielded a value of 575 for the RAR approach, versus 871 for PFA.

Figure 2 is inserted about here.

Next, we want to investigate the row and column interaction. This can be done by visual inspection of the biplot, as described in Gower and Hand (1996). Figure 3 shows the biplots based on the classical PFA (using the standard S-plus implementation) and on RAR.

Figure 3 is inserted about here.

The shape of the plots are quite different, as can be shown by a Procrustes analysis, and give rise to different interpretations. In the left plot, AL and TR are outlying in almost *all* variables, as are their values in Table 1. In the right plot, they are still visible as outliers, but the other points are much better represented. Take for example Hungary (H), which has extreme projections on almost all variables in classical biplot. But this is not corresponding well to the values in Table 1, whereas the presentation for Hungary in the robust biplot resembles quite well the real data values. The same exercise can be done for the other countries. To summarize, the biplot based on classical FA gives a good representation for AL and TR, but is also heavily influenced by them. Therefore, the representation of the other rows in the data matrix is rather poor. The RAR biplot gives an accurate representation of the big majority of the data, as we verified by computing the sum of squared differences between the observed and fitted values of the cells in the data matrix.

Note that in this example $n > p$, so it would be possible to start from a robust covariance matrix. However, in this example n is not very large relative to p. Even when using a maximal breakdown robust scatter estimator, this approach could break down if $4 = [(n - p + 1)/2]$ different rows contain an outlying cell (cfr. Davies 1987). Stated otherwise, we could already have breakdown if just 4 out of 144 cells in the data matrix are contaminated.

5.2 Artificial Data

As an example, an artificial data set of size $n = 50$ and $p = 7$ was generated according to the model (1.1) with $k = 2$. The factor scores were generated according to the standard bivariate normal $N_2(0, I_2)$. The bivariate loading parameters were generated so that they form three groups: $\lambda_1, \lambda_2, \lambda_3 \sim N_2((-1, 1)^\top, I_2/6)$, $\lambda_4, \lambda_5, \lambda_6 \sim N_2((-1, -1)^\top, I_2/6)$, and $\lambda_7 = (5, 0)^\top$. The uniqueness parameters are generated as $\psi_j \sim |N(1.5, 0.5)|$, and the errors

$\varepsilon_{ij} \sim N(0, \psi_j)$. Afterwards, 30 observations x_{ij} were replaced by severe outliers: we set $x_{ij} = 200$ for $(i = 21, \ldots, 35$ and $j = 6)$ and for $(i = 36, \ldots, 50$ and $j = 4)$. To these data we applied Principal Factor Analysis (using the classical correlation matrix) as well as the RAR method outlined in the previous section.

Figure 4 is inserted about here.

Figure 5 is inserted about here.

The robust biplot in Figure 4 reveals the true grouping of the loading vectors (the arrows), while the classical biplot fails to do so. The outliers are clearly visible in the robust residual plot, but not in the classical one (Figure 5). The example is of course artificial, but it shows that classical FA is not suitable as an outlier detection tool, as some practicioners believe. In fact, FA is not even intended to be a tool to detect outliers. But in many practical examples, some outliers (but not necessarily all of them) will show up in the classical biplot, since they attract the estimates of the scores and loadings towards them. This biplot will then however not give the factor structure of the data anymore, while this is the real purpose of a FA. We prefer the biplot to represent the true factor structure, and therefore we use a robust method. Outliers need not have an outlying projection in the true factor space, so they need not be visible in the robust biplot. In any case, the outliers will be visible in the robust residual plot.

6 Simulation

In this section we want to compare the performance of LS, L^1, RAR, MCD-based and classical principal factor analysis (PFA). We generate a matrix F of factor scores, with elements $f_{il} \sim N(0, 1)$ for $1 \leq i \leq n$ and $1 \leq l \leq k$. We also generate a matrix $\tilde{\Lambda}$ of loadings, with elements $\tilde{\lambda}_{jl} \sim U(-2, 2)$ (uniformly distributed in the range $[-2, 2]$) for $1 \leq j \leq p$ and $1 \leq l \leq k$. The unique variances $\tilde{\psi}_j$ are generated as $\tilde{\psi}_j \sim U(0, 1)$, and they are combined in the diagonal of the matrix $\tilde{\Psi}$. Furthermore, we generated a translation vector b with elements

$b_j \sim N(2, 10)$. These matrices and vectors together build a matrix X with elements

$$x_{ij} = \sum_{l=1}^{k} f_{il}\tilde{\lambda}_{jl} + b_j \tag{6.1}$$

for $1 \leq i \leq n$ and $1 \leq j \leq p$. We took $n = 20$, $p = 6$, and $k = 2$.

For $m = 1, \ldots, M = 4000$ simulations, we generated a noise term ε_{ij}^m distributed according to $N(0, \tilde{\psi}_j)$ to build the matrix $x_{ij}^m = x_{ij} + \varepsilon_{ij}^m$. But, for n_{out} entries, randomly placed in the data matrix, the noise term was generated from $N(0, 20)$, yielding up to n_{out} outlying cells. The number of outliers varied from $n_{out} = 0$ to 18, resulting in at most 15% of contaminated cells. It can be seen that the outliers in this example are not so severe. When having a look at the simulated data matrix X it would be difficult to pinpoint the outliers immediately.

The estimation procedure outlined in Section 2 was applied to the generated data sets x_{ij}^m. The standardization in Step 0 of the algorithm yields $\hat{b}_j^m$ and $\hat{s}_j^m$ as estimates of center and scale of the p variables. The center and the scale are estimated by the mean and the standard deviation for the non-robust methods LS and PFA, by the median and the MAD for the L^1 and RAR procedures, and by center and scale from the multivariate MCD estimator for the MCD-based FA.

Fitting the model gave estimated scores and loadings (computed from the standardized data), and allowed us to compute the fitted values

$$\hat{x}_{ij}^m = \hat{b}_j^m + \hat{s}_j^m \sum_{l=1}^{k} \hat{f}_{il}^m \hat{\lambda}_{jl}^m. \tag{6.2}$$

To measure the overall quality of the fit, we simulated the mean squared error by

$$\frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\hat{x}_{ij}^m - x_{ij} \right)^2 \right) \tag{6.3}$$

and the median squared error by

$$\frac{1}{M} \sum_{m=1}^{M} \mathrm{med}_{i,j} \left(\hat{x}_{ij}^m - x_{ij} \right)^2. \tag{6.4}$$

These measures are plotted in Figure 6 for different amounts of contamination, for the classical PFA, the MCD-based principal factor analysis, the principal component solution (2.2) to factor analysis (LS), the L^1 fit (2.3), and the RAR estimator.

When there is no contamination, the classical procedures have the smallest mean/median squared error. The LS-based estimator is optimal by construction, but it looses this optimality even in presence of very small percentages of outliers. In presence of outliers, the

RAR estimator is outperforming all other considered estimators with respect to the above defined measures. In particular, we see that it is necessary to take the weighted L_1 procedure instead of using ordinary L_1.

Figure 6 is inserted about here.

Figure 7 is inserted about here.

To compute the empirical efficiency of the estimators, we need to take into account that we work with standardized data. The population covariance matrix equals $\Sigma = \tilde{\Lambda}\tilde{\Lambda}^\top + \tilde{\Psi}$, which can be rewritten as $R = \Lambda\Lambda^\top + \Psi$, with R the population correlation matrix, $\Lambda = D_\Sigma^{-1/2}\tilde{\Lambda}$ and $\Psi = D_\Sigma^{-1/2}\tilde{\Psi}D_\Sigma^{-1/2}$, with $D_\Sigma = \mathrm{diag}(\Sigma)$. The reduced correlation matrix $A = \Lambda\Lambda^\top$ is then estimated by $\hat{a}_{ij}^m = \sum_{l=1}^k \hat{\lambda}_{il}^m \hat{\lambda}_{jl}^m$ for $m = 1, \ldots, M = 4000$. Note that we need to do the orthogonalization (Step 3 of the RAR algorithm) here, in order to have a uniquely identified reduced correlation matrix. Since the loadings are not uniquely determined, we focus on the estimation of the reduced correlation matrix. The precision of the estimator of the reduced correlation matrix is measured by the following mean squared error (MSE):

$$\frac{1}{M}\sum_{m=1}^{M}\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\hat{a}_{ij}^m - a_{ij}\right)^2. \tag{6.5}$$

In Figure 7a we see how this MSE varies with the amount of contamination, for the different estimators. Finally, the MSE of the uniquenesses is computed as

$$\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p}\sum_{j=1}^{p}\left(\hat{\psi}_j^m - \psi_j\right)^2 \tag{6.6}$$

and shown in Figure 7b. We see that for smaller amounts of contamination MCD performs the best, closely followed by RAR. But for larger amounts of contamination ($\geq 10\%$) it is again the RAR procedure which is more accurate. It is remarkable to see that the LS method yields more precise estimates than the PFA method for the parameters of the factor model (in presence of contamination), despite of the fact that the latter method exploits the presence of the specific variances.

Since the number of replications of the simulation is quite large ($M = 4000$), the standard errors of the measures (6.3) to (6.6) are very small. Here they were all smaller than 0.04, and often much smaller.[1] As a conclusion of this simulation experiment, we can say that it favours the RAR estimator.

7 Applying RAR to the FANOVA Model

The standard model for a two-way table is the ANOVA model

$$x_{ij} = \mu + a_i + b_j + \delta_{ij} \tag{7.1}$$

where μ is called the overall mean, a_i represents the row effect and b_j the column effect. In a classical setup, the row and column effects are assumed to have zero mean. The terms δ_{ij} can either be seen as residuals or as interaction terms between rows and columns. Expression (7.1) is called an additive model. It is however quite possible that the interaction terms δ_{ij} still contain some structure that can be described by a factor model $\delta_{ij} = \sum_{l=1}^{k} \lambda_{jl} f_{il} + \varepsilon_{ij}$ as in (1.1), yielding the overall model

$$x_{ij} = \mu + a_i + b_j + f_i^\top \lambda_j + \varepsilon_{ij}. \tag{7.2}$$

This is the FANOVA model (cfr. Gollob 1968, Denis and Gower 1996, and the references therein), which combines aspects of analysis of variance and factor analysis. Among others, Gabriel (1978) considered models like (7.2) and estimated the unknown parameters using a least squares fit. A first idea would be to proceed sequentially by estimating the additive model first, and afterwards performing a factor analysis on the residuals. But better fits can be obtained by estimating all parameters jointly. For the least squares fit there is no difference between the simultaneous and the sequential approach, but this is no longer true for the robust fits. Therefore we will estimate additive and multiplicative terms simultaneously.

The RAR estimator for the FANOVA model can be defined as in Section 2. Denote by θ the vector collecting the scores, loadings, row and column effects and the overall effect μ. In order to estimate the $(k+1)(n+p)+1$ unknown elements of θ from the np available data values, we can use the RAR estimator defined as in (2.6):

$$\hat{\theta}_{RAR} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{p} w_i(\theta) v_j(\theta) |x_{ij} - \hat{x}_{ij}(\theta)| \tag{7.3}$$

[1] An exception is the mean squared error of the overall fit for the MCD-based method where the standard error increases to 0.36 for higher contamination.

with $\hat{x}_{ij}(\theta) = \mu + a_i + b_j + f_i^\top \lambda_j$. The weights w_i and v_j are defined as in (2.7), and are downweighting outlying scores and loadings in the k-dimensional spaces of scores and loadings. To uniquely identify the parameters in (7.2), the function (7.3) will be minimized under the constraints

$$\operatorname*{med}_i(a_i) = \operatorname*{med}_j(b_j) = 0 \quad \text{and} \quad \operatorname*{med}_i(f_{il}) = \operatorname*{med}_j(\lambda_{jl}) = 0 \tag{7.4}$$

for $l = 1, \ldots, k$. The constraints (7.4) are consistent with a robust approach. The algorithm to compute the RAR estimator in FANOVA models is based on alternating regressions, and is almost identical to the iterative scheme outlined in Section 2. One difference is that, instead of working with regression through the origin, intercepts need to be estimated as well. (The S-Plus program for applying RAR to FANOVA models can be retrieved from the website mentioned before.) In the simplified model (7.1), the RAR approach coincides with the median polish technique (see Hoaglin, Mosteller and Tukey 1983).

As an example, we consider the logarithm of the real income per capita in 18 European countries from 1962 to 1994, as obtained from EUROSTAT. (More precisely, this is the gross domestic product (GDP), deflated by the GDP deflater to get 1990 market prices, and divided by the population.) Instead of representing the data in an 18 by 33 matrix, Figure 8 plots each row of the data matrix x_{ij} as a time series.

Figure 8 is inserted about here.

One sees that there is an upward tendency in each time series, and that some countries have higher income/capita than others.

We now fit the FANOVA model by means of RAR. Figure 9a shows the row effects $\hat{a}_i$, which are country effects: they indicate the deviation of the median level of the ith time sequence x_{ij} from the overall median μ. We observe the highest median level for Switzerland (CH), while Greece (GR) and Portugal (P) have the lowest income levels. The time effects $\hat{b}_j$ are plotted in Figure 9b as a time series smoothed by the LOESS method of Cleveland (1979).

Figure 9 is inserted about here.

We see an increasing trend, corresponding to economic growth in the studied period. Now denote by $y_{ij} = \exp(x_{ij})$ the untransformed data. Neglecting the error term for a moment, the first difference of the sequence of time effects equals

$$\Delta b_j = b_j - b_{j-1} = \log \frac{\text{med}_i\, y_{ij}}{\text{med}_i\, y_{i,j-1}} \approx \frac{\text{med}_i\, y_{ij} - \text{med}_i\, y_{i,j-1}}{\text{med}_i\, y_{i,j-1}}. \tag{7.5}$$

This corresponds to the (relative) growth rate of the median income level over the different countries. (The approximation $\approx$ in (7.5) is due to a first order Taylor expansion.) If we used the purely additive model (7.1) then we would believe that the expected growths for all individual countries are the same, and equal to Δb_j. This would imply that the data in Figure 8 could be modeled by a collection of parallel curves, all equal to $\mu + b_j$ plus a constant shift term a_i. This is clearly not the case. The FANOVA model (7.2) allows to go beyond the hypothesis of parallel curves, while still remaining parsimonious. (For a similar reason, but in another context, the FANOVA model was used by Gauch 1988.)

To select the number of factors for the FANOVA model, a measure analogous to (4.1) has been computed for different values of k. The value $R^2_{RAR}(k)$ measures how much more variability is explained by adding k factor terms to the purely additive model:

$$R^2_{RAR}(k) = 1 - \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{p} w_i v_j |x_{ij} - \hat{x}_{ij}|}{\sum_{i=1}^{n} \sum_{j=1}^{p} w_i v_j |x_{ij} - \hat{x}^0_{ij}|} \right)^2 \tag{7.6}$$

where $\hat{x}^0_{ij}$ is the purely additive fit to the data matrix. The obtained values are

k	1	2	3
$R^2_{RAR}(k)$	0.58	0.80	0.90

indicating that it is quite reasonable to model the interaction terms in (7.3) with two factors. In Figure 10b the estimated loadings $\hat{\lambda}_j$ are pictured as a time sequence. While Figure 9b summarized all 18 time series in a single one, Figure 10b gives information about secondary features of the data. The first sequence of loadings $\hat{\lambda}_{j1}$ has an increasing trend and is almost linear, as was the main time effect. Countries with high values for the first factor will therefore have a larger slope, and hence a faster growth rate. We see from Figure 10a that Luxemburg has a quite large growth rate over the period in question, as opposed to Switzerland (see also Figure 8). The second series of loadings $\hat{\lambda}_{j2}$ can be interpreted as the impact of the global macro-economic evolution: it increases up to 1973 (the oil crisis) and then goes into a period of recession until the mid-eighties. This second factor corresponds with our subjective feeling of the evolution of our incomes. In Figure 10a we see that Belgium, France and West-Germany are close to the center of the plot, indicating that they

are representative for the evolution of incomes in the 18 countries (not for the absolute levels, which are captured by the country effects). Greece and Portugal are outlying for the second factor, and in Figure 8 we indeed see that their growth rate decreased significantly after 1973.

Figure 10 is inserted about here.

---

8 Conclusions

Many classical techniques of multivariate statistics are based on the sample covariance matrix $\hat{\Sigma}$. Since $\hat{\Sigma}$ is very sensitive to outliers, the resulting methods are not robust. One way to robustify these procedures is to insert a robust covariance matrix instead, as was done in the context of principal components (e.g. by Devlin et al 1981, Croux and Haesbroeck 2000), canonical correlations (Croux and Dehon 2002), canonical variates (Campbell 1982) biplots (Daigle and Rivest 1992) and many other papers (e.g. Visuri et al 2000). In this paper we propose the RAR method for factor analysis, which works well for both $p < n$ and $p \geq n$. We stress that in many applications $p \geq n$, for instance in chemometrics, and that robust statistical methods are needed. The price we pay for this general applicability is a longer computation time.

We believe that RAR has many virtues as an estimator and as a data analytic tool. In the simulation experiment in Section 6, the quality of the fit of the lower rank matrix $\hat{X}$ to the data matrix using RAR was shown to be superior. This implies that for the construction of robust biplots the RAR approach is preferable.

Another advantage of the RAR method is that it can withstand a higher number of outlying cells than FA based on robust scatter matrix estimators. The approach based on an $\alpha\%$ breakdown scatter matrix estimator and the RAR approach based on an $\alpha\%$ breakdown regression estimator have the same theoretical breakdown value $\alpha\%$ for the estimation of loadings and specific variances, but RAR is more robust in practice. Indeed, if a row (case) has an outlying cell (coordinate), the robust scatter matrix estimator will declare the entire row as outlying, and it will not try to fit the other cells of that row. The RAR estimator will still use the information in those other cells. When the contaminated rows have their outlying cells in different columns, RAR can withstand more outlying rows than the robust

scatter approach can. This is analogous to the treatment of missing data in data tables. A missing cell value should not necessarily imply deletion of all the other cells in that row. Therefore, the RAR method could also be used for performing factor analysis on data with missing values.

Acknowledgments: We wish to thank the referees for interesting and helpful comments. Moreover, we are grateful to Antoine de Falguerolles for stimulating discussions.

References

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

Becker C. and Gather U. 2001. The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules, Computational Statistics and Data Analysis, 36:119-127.

Bloomfield P. and Steiger W.L. 1983. Least Absolute Deviations: Theory, Applications, and Algorithms, Boston, Mass, Birkhäuser.

Campbell N.A. 1982. Robust procedures in multivariate analysis II: Robust canonical variate analysis, Applied Statistics, 31:1-8.

Cleveland W.S. 1979. Robust locally weighted regression and smoothing scatter plots, Journal of the American Statistical Association, 74:829-836.

Croux C. and Dehon C. 2002. Analyse canonique basee sur des estimateurs robustes de la matrice de covariance, To appear in La Revue de Statistique Appliquee.

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.

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 Ruiz-Gazen A. 1996. A fast algorithm for robust principal components based on projection pursuit. COMPSTAT 1996, Proceedings in Computational Statistics (ed. A. Prat), Heidelberg, Physica-Verlag, pp. 211-216.

Daigle G. and Rivest L.-P. 1992. A robust biplot, The Canadian Journal of Statistics, 20:241-255.

Davies L. 1987. Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices, The Annals of Statistics, 15:1269-1292.

de Falguerolles A. and Francis B. 1992. Algorithmic approaches for fitting bilinear models. COMPSTAT 1992, Proceedings in Computational Statistics, Vol. 1 (eds. Y. Dodge and J. Whittaker), Heidelberg, Physica-Verlag, pp. 77-82.

Denis J.-B. and Gower J.C. 1996. Asymptotic confidence regions for biadditive models: Interpreting genotype-environment interactions, Applied Statistics, 45:479-493.

Devlin S.J., Gnanadesikan R., and Kettenring J.R. 1981. Robust estimation of dispersion matrices and principal components, Journal of the American Statistical Association, 76:354-362.

El Bantli F. and Hallin M. 1999. L1-estimation in linear models with heterogeneous white noise, Statistics and Probability Letters, 45:305-315.

Filzmoser P. 1999. Robust principal components and factor analysis in the geostatistical treatment of environmental data, Environmetrics, 10:363-375.

Gabriel K.R. 1978. Least squares approximation of matrices by additive and multiplicative models, Journal of the Royal Statistical Society B, 40(2):186-196.

Gabriel K.R. 1998. Generalized bilinear regression, Biometrika, 85:689-700.

Gabriel K.R. and Zamir S. 1979. Lower rank approximation of matrices by least squares with any choice of weights, Technometrics, 21:489-498.

Gauch H.G. 1988. Model selection and validation for yield trial with interaction, Biometrics, 44:705-716.

Gifi A. 1990. Nonlinear Multivariate Analysis, Chichester, Wiley & Sons.

Gollob H.F. 1968. A statistical model which combines features of factor analytic and analysis of variance techniques, Psychometrika, 33:73-116.

Gower J. and Hand D. 1996. Biplots, New York, Chapman & Hall.

Hallin M. and Mizera I. 2001. Sample heterogeneity and M-estimation, Journal of Statistical Planning and Inference, 93:139-160.

Hoaglin D., Mosteller F., and Tukey J. 1983. Understanding Robust and Exploratory Data Analysis, New York, Wiley & Sons.

Johnson R.A. and Wichern D.W. 1998. Applied Multivariate Statistical Analysis, 4th edition, New Yersey, Prentice Hall.

Kosfeld R. 1996. Robust exploratory factor analysis, Statistical Papers, 37:105-122.

Li G. and Chen Z. 1985. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo, Journal of the American Statistical Association, 80:759-766.

Martens H. and Naes T. 1989. Multivariate Calibration, New York, Wiley & Sons.

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

Rousseeuw P.J. 1985. Multivariate estimation with high breakdown point. Mathematical Statistics and Applications, Vol. B (eds. W. Grossmann et al, Dordrecht, Reidel, pp. 283-297.

Rousseeuw P.J. and van Zomeren B.C. 1990. Unmasking multivariate outliers and leverage points, Journal of the American Statistical Association, 85:633-639.

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

Simpson D.G., Ruppert D., and Carroll R.J. 1992. On one-step GM estimates and stability of inferences in linear regression, Journal of the American Statistical Association, 87:439-450.

Tanaka Y. and Odaka Y. 1989a. Influential observations in principal factor analysis, Psychometrika, 54(3):475-485.

Tanaka Y. and Odaka Y. 1989b. Sensitivity analysis in maximum likelihood factor analysis, Communications in Statistics–Theory and Methods, A18(11):4067-4084.

Terbeck W. and Davies P. 1998. Interactions and outliers in the two-way analysis of variance, The Annals of Statistics, 26:1279-1305.

Ukkelberg Å. and Borgen O. 1993. Outlier detection by robust alternating regressions, Analytica Chimica Acta, 277:489-494.

Visuri S., Koivunen V., and Oja H. 2000. Sign and rank covariance matrices, Journal of Statistical Planning and Inference, 91:557-575.

Wold H. 1966. Nonlinear estimation by iterative least squares procedures. Research Papers in Statistics: Festschrift for Jerzy Neyman (ed. F.N. David), New York, John Wiley, pp. 411-444.

Table Captions

Table1: European health and fertility data. The 16 countries are Austria (A), Albania (AL), Bulgaria (BG), Switzerland (CH), Czechoslovakia (CS), East Germany (DDR), Hungary (H), Norway (N), Poland (PL), Rumania (RO), Sweden (S), Finland (SF), Soviet Union (SU), Turkey (TR), Yugoslavia (YU), and the European Community (EU).

Figure Captions

Figure 1: R^2-plot and scree R^2-plot of the European population data for LS, L^1, and the RAR estimator.

Figure 2: Residual plot of classical FA (left) and RAR (right) for the European health and fertility data.

Figure 3: Biplot of the European health and fertility data, obtained from classical principal factor analysis (left) and from RAR (right).

Figure 4: Biplot of the artificial data for classical FA (left) and RAR (right).

Figure 5: Artificial data: residual plot of classical FA (left) and RAR (right).

Figure 6: Quality of the fits under contamination, using (a) the mean squared error criterion, and (b) the median squared error criterion.

Figure 7: MSE of the estimates of (a) the reduced rank correlation matrix, and (b) the uniquenesses, under various levels of contamination.

Figure 8: Log Real income per capita of 18 European countries in the years 1962-1994. Countries are Belgium (B), Denmark (DK), West Germany (D), Greece (GR), Spain (E), France (F), Ireland (IRL), Italy (I), Luxembourg (L), Netherlands (NL), Portugal (P), United Kingdom (GB), Switzerland (CH), Austria (A), Norway (N), Sweden (S), Finland (SF), and Iceland (IS). To avoid overplotting, only six labels are shown here.

Figure 9: Estimated (a) country and (b) time effects of the income/capita data analyzed with RAR.

Figure 10: Estimated (a) scores and (b) loadings for a two-factor FANOVA model for the income data, obtained with RAR. The solid line is a smooth fit for the loadings of factor 1 and the dotted line for the loadings of factor 2.

Table 1: European health and fertility data. The 16 countries are Austria (A), Albania (AL), Bulgaria (BG), Switzerland (CH), Czechoslovakia (CS), East Germany (DDR), Hungary (H), Norway (N), Poland (PL), Rumania (RO), Sweden (S), Finland (SF), Soviet Union (SU), Turkey (TR), Yugoslavia (YU), and the European Community (EU).

	pop_growth	give_birth	women%	lifeexp_f	lifeexp_m	inf_mort	inhab/doc	calorie	baby_underw
A	-0.1	48	110	77	70	10	440	3440	6
AL	1.8	50	97	75	68	41	2100	2716	7
BG	0.2	47	101	75	69	15	400	3593	6
CH	0.0	44	103	80	74	7	390	3406	5
CS	0.3	46	105	75	66	14	350	3473	6
DDR	0.0	47	110	75	68	9	490	3769	6
H	-0.1	46	106	75	67	19	390	3544	10
N	0.2	48	101	80	74	9	460	3171	4
PL	0.6	48	104	76	68	18	550	3224	8
RO	0.5	47	102	73	68	26	700	3413	6
S	0.0	47	101	80	74	6	410	3007	4
SF	0.2	47	107	79	72	6	460	2961	4
SU	0.7	48	112	73	64	30	270	3332	6
TR	1.9	49	97	67	62	79	1530	3218	8
YU	0.5	51	103	74	68	27	700	3499	7
EU	0.2	48	104	78	73	10	509	3421	5

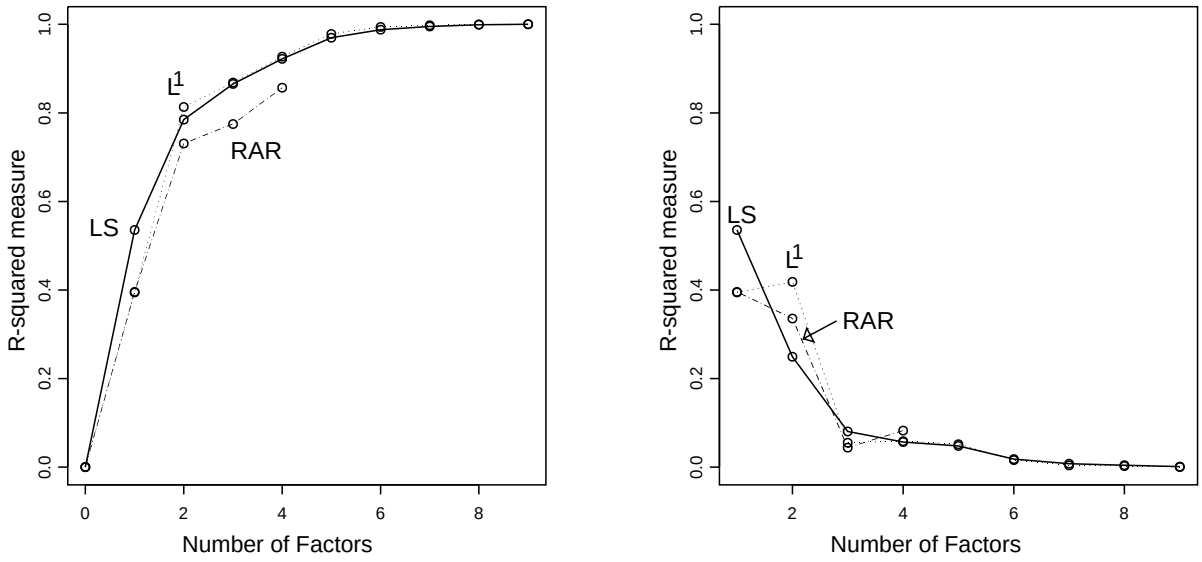

Figure 1: R^2-plot and scree R^2-plot of the European population data for LS, L^1, and the RAR estimator.

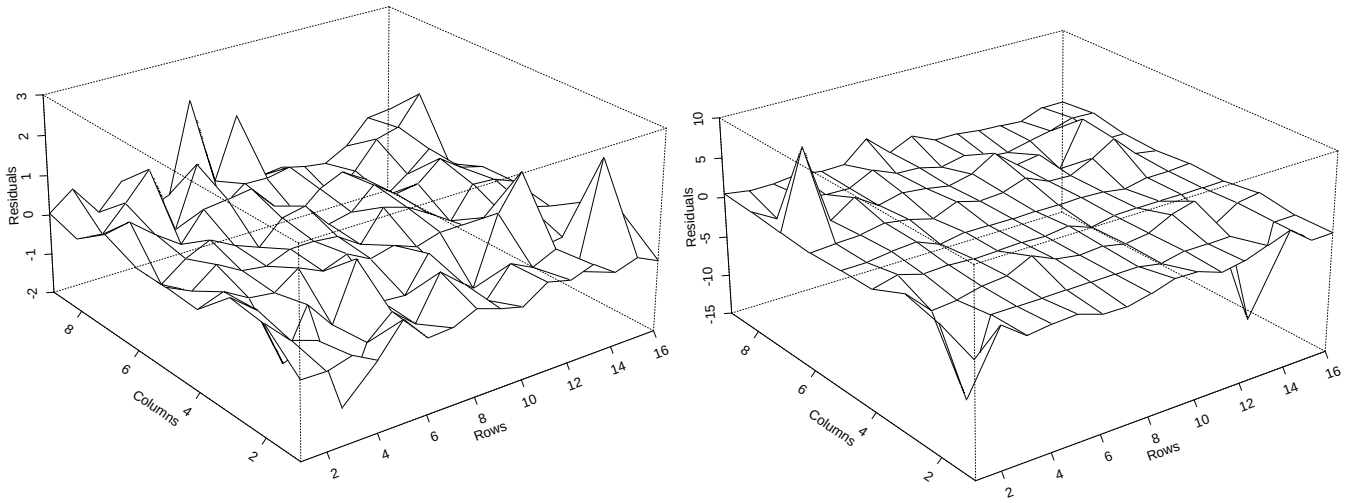

Figure 2: Residual plot of classical FA (left) and RAR (right) for the European health and fertility data.

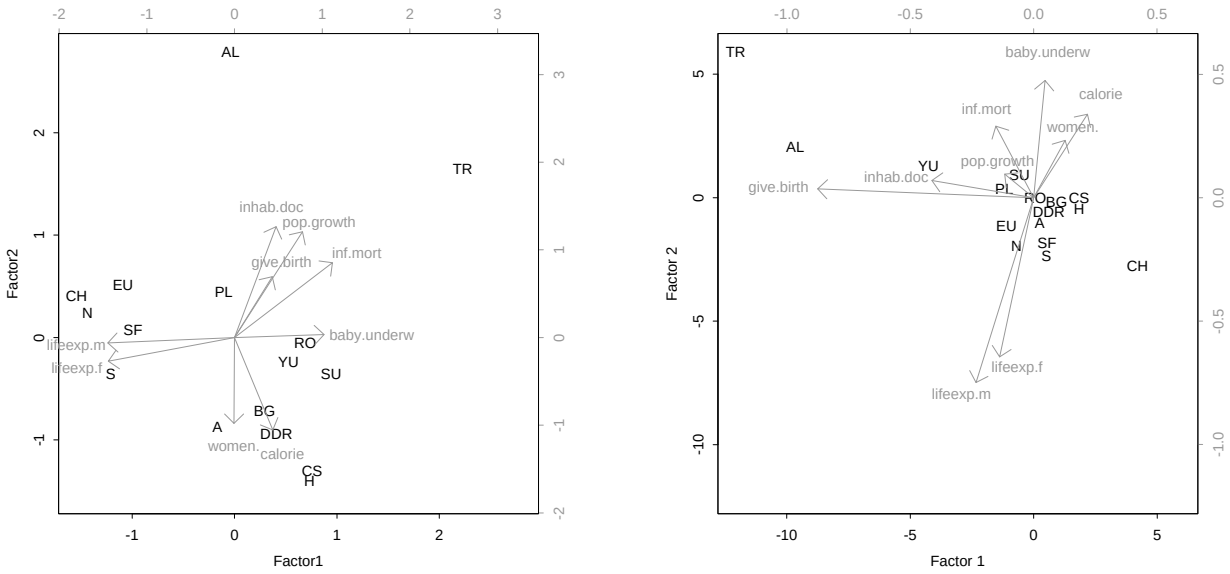

Figure 3: Biplot of the European health and fertility data, obtained from classical principal factor analysis (left) and from RAR (right).

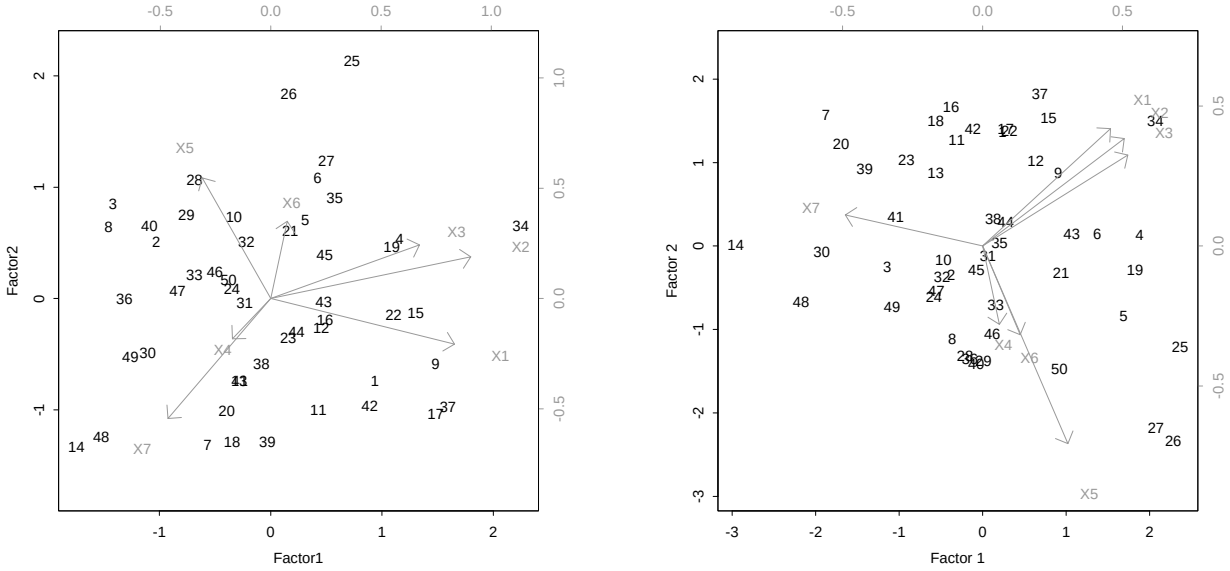

Figure 4: Biplot of the artificial data for classical FA (left) and RAR (right).

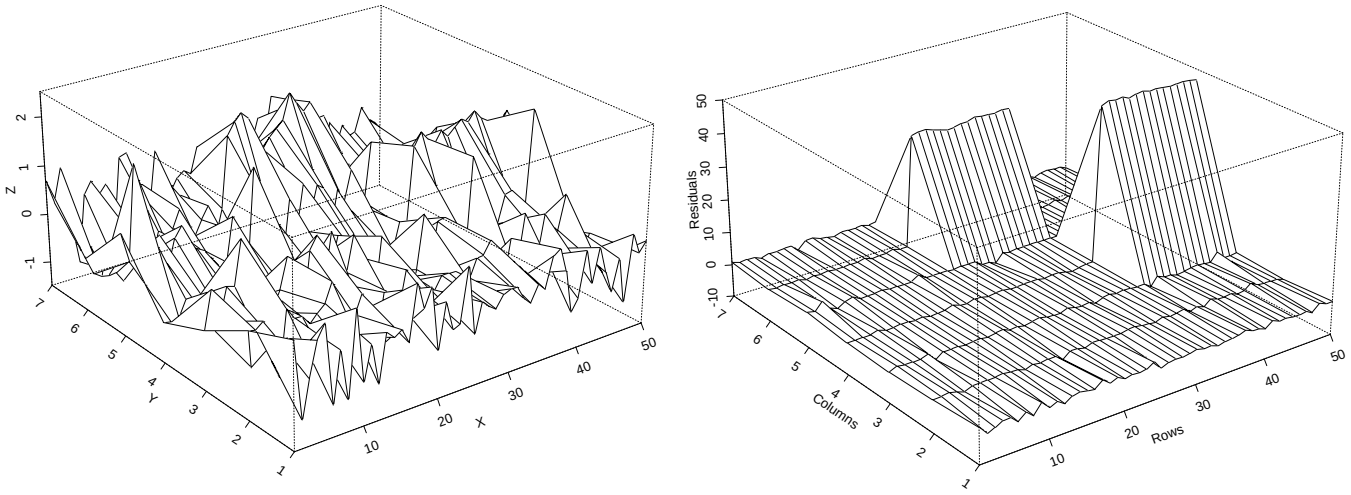

Figure 5: Artificial data: residual plot of classical FA (left) and RAR (right).

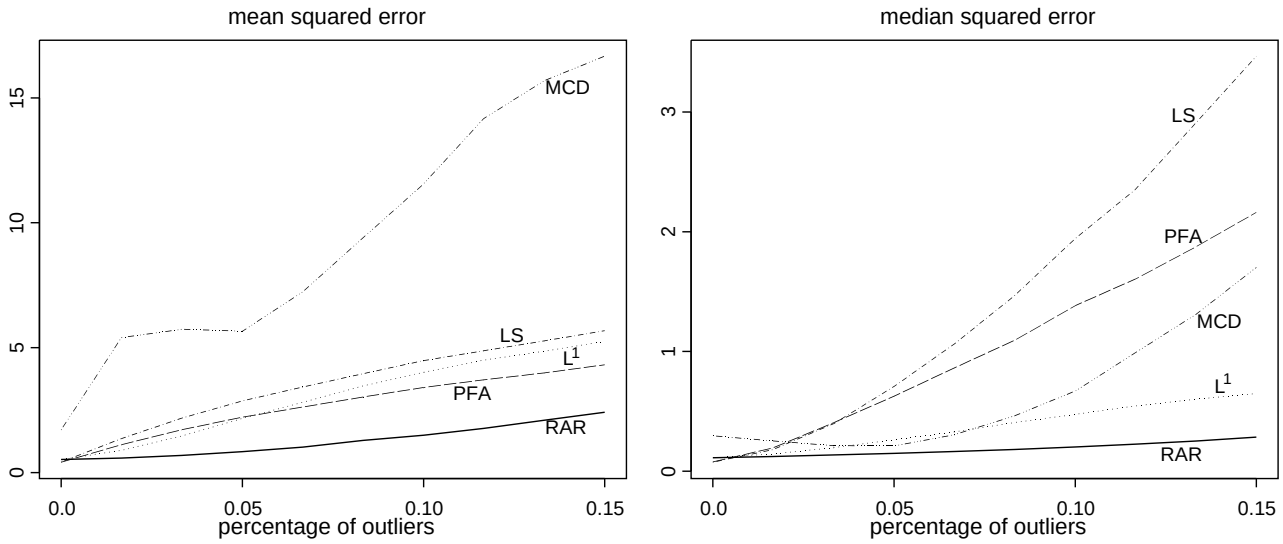

Figure 6: Quality of the fits under contamination, using (a) the mean squared error criterion, and (b) the median squared error criterion.

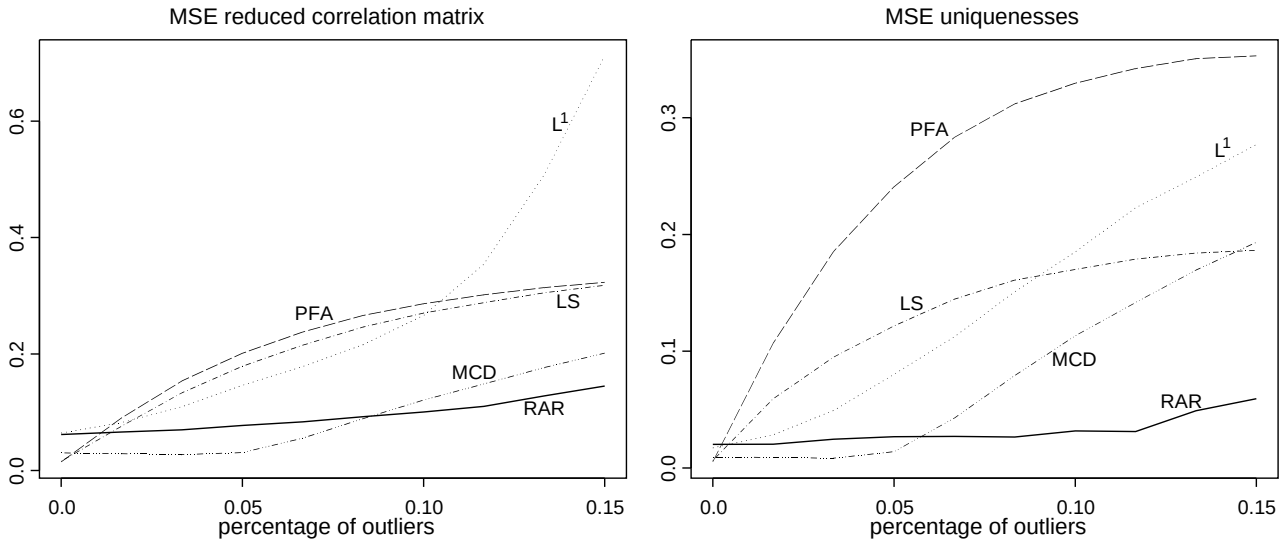

Figure 7: MSE of the estimates of (a) the reduced rank correlation matrix, and (b) the uniquenesses, under various levels of contamination.

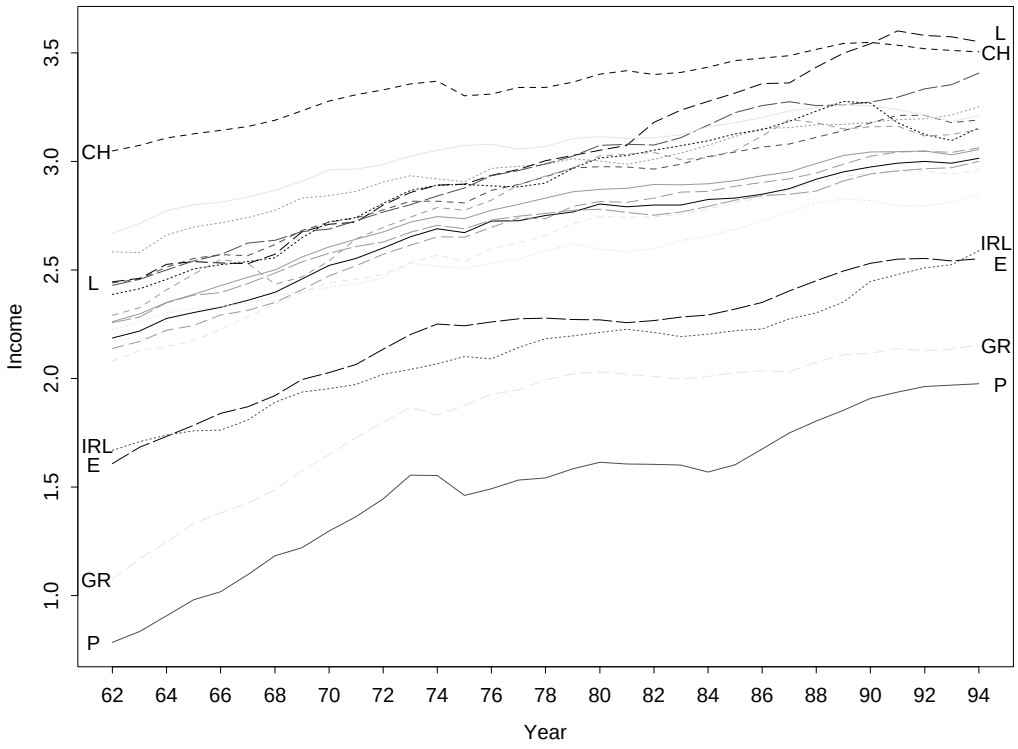

Figure 8: Log Real income per capita of 18 European countries in the years 1962-1994. Countries are Belgium (B), Denmark (DK), West Germany (D), Greece (GR), Spain (E), France (F), Ireland (IRL), Italy (I), Luxembourg (L), Netherlands (NL), Portugal (P), United Kingdom (GB), Switzerland (CH), Austria (A), Norway (N), Sweden (S), Finland (SF), and Iceland (IS). To avoid overplotting, only six labels are shown here.

Country effect $\hat{a}_i$ Time effect $\hat{b}_j$

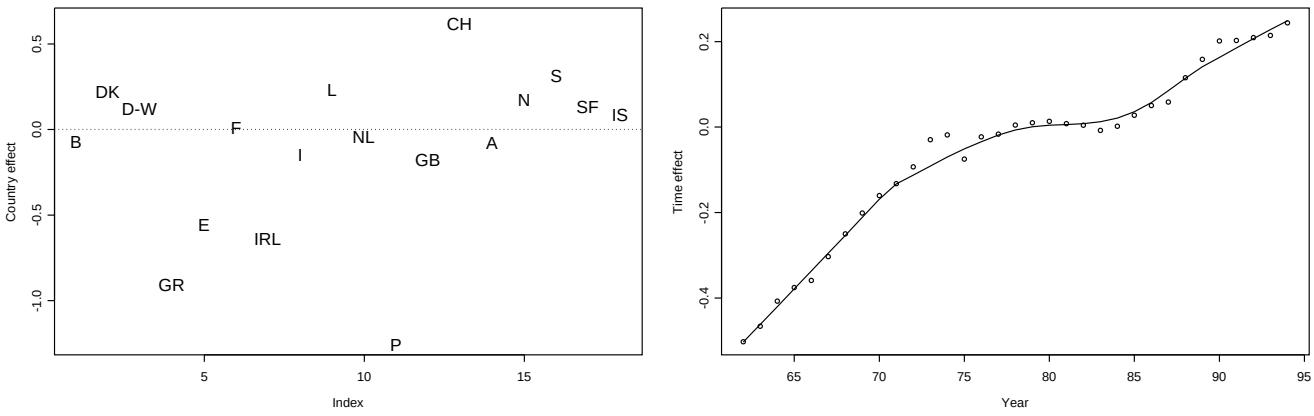

Figure 9: Estimated (a) country and (b) time effects of the income/capita data analyzed with RAR.

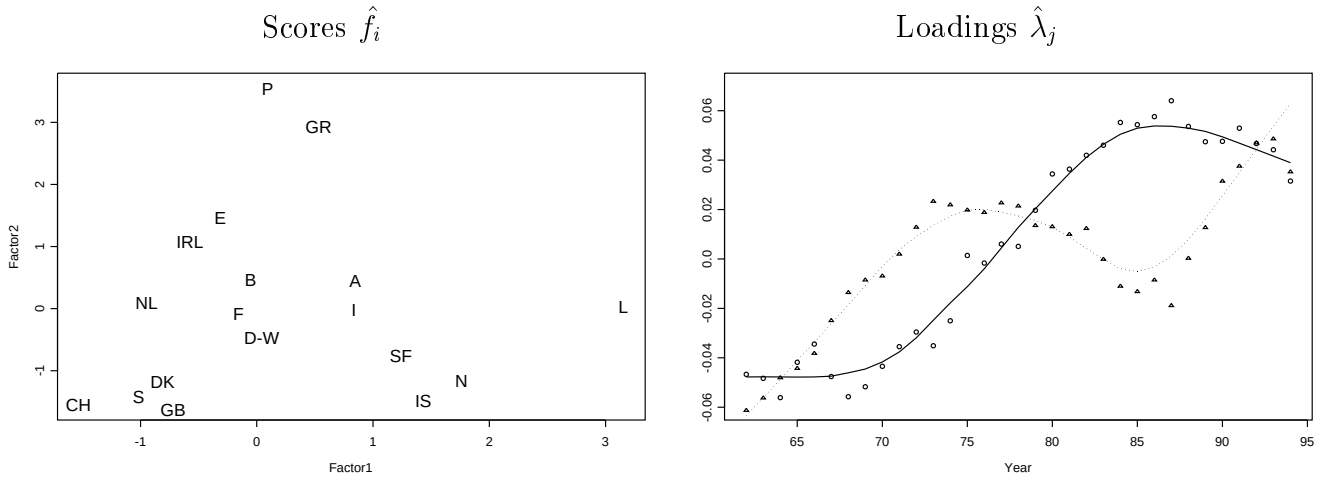

Figure 10: Estimated (a) scores and (b) loadings for a two-factor FANOVA model for the income data, obtained with RAR. The solid line is a smooth fit for the loadings of factor 1 and the dotted line for the loadings of factor 2.