Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

# Detecting Outliers in Factor Analysis Using the Forward Search Algorithm

Dimitris Mavridis [a] & Irini Moustaki [b]
[a] The School of Mathematics, University of Edinburgh,
[b] Department of Statistics, London School of Economics,

Available online: 10 Sep 2008

To cite this article: Dimitris Mavridis & Irini Moustaki (2008): Detecting Outliers in Factor Analysis Using the Forward Search Algorithm, Multivariate Behavioral Research, 43:3, 453-475

To link to this article: http://dx.doi.org/10.1080/00273170802285909

*Multivariate Behavioral Research*, 43:453–475, 2008

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273170802285909

Psychology Press
Taylor & Francis Group



# Detecting Outliers in Factor Analysis Using the Forward Search Algorithm

Dimitris Mavridis
*The School of Mathematics*
*University of Edinburgh*

Irini Moustaki
*Department of Statistics*
*London School of Economics*

In this article we extend and implement the forward search algorithm for identifying atypical subjects/observations in factor analysis models. The forward search has been mainly developed for detecting aberrant observations in regression models (Atkinson, 1994) and in multivariate methods such as cluster and discriminant analysis (Atkinson, Riani, & Cerioli, 2004).

Three data sets and a simulation study are used to illustrate the performance of the forward search algorithm in detecting atypical and influential cases in factor analysis models. The first data set has been discussed in the literature for the detection of outliers and influential cases and refers to the grades of students on 5 exams. The second data set is artificially constructed to include a cluster of contaminated observations. The third data set measures car's characteristics and is used to illustrate the performance of the forward search when the wrong model is specified. Finally, a simulation study is conducted to assess various aspects of the forward search algorithm.

Factor analysis is a widely used statistical technique that aims to explain the inter-relationships among a set of $p$ observed variables by $q$ latent variables

Correspondence concerning this article should be addressed to Irini Moustaki, London School of Economics, Houghton Street, London WC2A 2AE, UK. E-mail: I.Moustaki@lse.ac.uk

(factors) where $q$ is much smaller than $p$. Some key references are Jöreskog (1967), Lawley and Maxwell (1971), and Bartholomew and Knott (1999).

Often, outliers are errors occurring with data recording that can usually be identified by simple variable and cross-variable checking. Here, we are interested in detecting observations that are most unlikely to occur under the hypothesized model.

It has been noted that the presence of outliers in the data may distort both the estimated model parameters and the goodness-of-fit of the model. Bollen and Arminger (1991) argue that a single outlier may be responsible for the addition of an extra factor in the model and Bollen (1989) argues that outliers may be responsible for negative error variances known also as Heywood cases.

Factor analysis fits the model to the sample covariance matrix. Yuan and Bentler (1998b) pointed out that the sample covariance matrix has unbounded influence function and zero breakdown point and therefore it is unlikely to yield reliable results even in the presence of a single outlier. Jöreskog (1979) suggests that the factor model should be fitted to a robust covariance matrix instead on the grounds that the normality assumption does not necessarily hold when outliers are present in the data.

There are many papers that discuss alternative ways of obtaining a robust location and dispersion measure for multivariate continuous variables (Campbell, 1980; Cheng & Victoria-Feser, 2002; Davies, 1987; Devlin, Gnanadesikan, & Kettenring, 1981; Donoho, 1982; Hadi, 1992; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Maronna, 1976; Rousseeuw, 1985; Rousseeuw & Leroy, 1987; Rousseeuw & Van Zomeren, 1990; Stahel, 1981; Woodruff & Rocke, 1949) as well as papers on robust factor analysis (Filzmoser, 1999; Moustaki & Victoria-Feser, 2006; Pison, Rousseeuw, Filzmoser, & Croux, 2003; Yuan & Bentler, 1998a, 1998b, 2001). In all those approaches, outlying cases are being down-weighted so that they have minimum influence on the estimated means and covariance matrix or directly on the parameter estimates of the factor model.

The aim of this article is not to estimate another type of robust covariance matrix that will be used in the factor analysis model but instead to develop model diagnostics that will reveal the effect that each observation has on model parameter estimates, on goodness-of-fit statistics and fit measures and consequently on model inference. Detection of outliers in factor models has not received much attention primarily because of the latent nature of the factors. In regression analysis, deletion diagnostics are computed for measuring the exact effect of the deletion of a single observation on parameter estimates, $t$ tests, and residuals (prediction error). Some of the well-known deletion diagnostics are the deletion residuals and Cook's statistic (Cook & Weisberg, 1982). Those are known as "backward" methods because they start by fitting the model to the whole data set and then they delete one observation at a time. Here we discuss the forward

search (FS) algorithm that was initially developed for regression models. The FS is an iterative procedure that aims to provide an ordering of the data by their closeness to the fitted model and to use plots for identifying aberrant observations. The FS starts by fitting the model to a small subset of the whole data set and proceeds until all observations are included. Applications of the FS can be found in Atkinson and Riani (2000) and Atkinson et al. (2004). Here, we extend the FS algorithm to detecting outliers in factor analysis.

The article is organized as follows: the factor model is briefly discussed in the next section, followed by the steps of the forward search algorithm in factor analysis. The performance of the method is illustrated with real and simulated data. Finally, the main findings of the article are summarized.



## FACTOR ANALYSIS MODEL

We present here briefly the normal linear factor analysis model (Jöreskog, 1967; Lawley & Maxwell, 1971). Consider that we have $p$ manifest or observed variables also known as items and $q$ latent or unobserved variables also known as factors. The vector of observed variables is denoted by $\mathbf{y}$ where $\mathbf{y}' = (y_1, y_2, ..., y_p)$ and the vector of latent variables will be denoted by $\mathbf{z}$ where $\mathbf{z}' = (z_1, z_2, ..., z_q)$. Both manifest and latent variables are assumed to be continuous. The Normal Linear Factor Model (NLFM) is

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda z} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\mu}$ is a $p \times 1$ vector of mean values for the manifest variables, $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, and $\boldsymbol{\epsilon}$ is a $p \times 1$ vector of error terms also known as *specific* or *unique factors* because they are unique to a particular $y_i$. The latent variables are taken to have a multivariate normal distribution, $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{\Phi})$, where $\boldsymbol{\Phi}$ denotes the covariance matrix of the latent variables. Furthermore, it is assumed that $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \boldsymbol{\Psi})$ and $E(\boldsymbol{z\epsilon}') = 0$, where $\boldsymbol{\Psi}$ is a diagonal matrix with elements the variances of the error terms also known as specific or unique variances. It follows that the joint distribution of the observed variables is a multivariate normal:

$$\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Lambda \Phi \Lambda}' + \boldsymbol{\Psi}).$$

The factor analysis model can be estimated either with maximum likelihood (Jöreskog, 1967) or with estimation methods that do not require the normality assumption such as the unweighted least squares estimation methods (ULS) and the generalized least squares (Jöreskog, 1972).

## FORWARD SEARCH

The forward search starts with a small subset of the data that is intended to be outlier-free and proceeds by adding observations until all are included. The steps of an FS can be summarized as follows:

- Choose an outlier-free initial subset of size $n_g$ from the $\binom{n}{n_g}$ possible subsets where $n$ is the total sample size. Those $n_g$ observations constitute the initial subset. This is the "basic" set formed at the beginning of the search whereas the remaining $(n - n_g)$ observations constitute the "non-basic" set.
- Find ways to progress in the FS so that eventually all observations from the "non-basic" set are included in the "basic" set.
- Monitor statistics of interest such as parameter estimates and goodness-of-fit tests and fit measures during the progress of the search.

Because we start with an initial sample of $n_g$ observations, there are $n - n_g$ iterations until all observations are included. The "basic" and the "non-basic" sets are mutually exclusive throughout the search and their union gives at each iteration the sample space. We use $\boldsymbol{Y}_l^{bs}$ to denote those subjects that are in the "basic" set at iteration $l$ and $\boldsymbol{Y}_l^{nbs}$ to denote those subjects that are in the "non-basic" set.

We now explain in detail the three steps of the FS in relation to the factor analysis model.

### Choosing the Initial Subset for the Forward Search

The size of the initial subset must be defined first. That size can be as small as $\frac{1}{2}p(p+1)+1$. This is the number of distinct elements in the variance-covariance matrix of $\mathbf{Y}$ plus one. A large initial subset will give more stable parameter estimates and smoother forward plots, increasing at the same time the chance of starting the search from a non-outlier-free subset. In practice, it is impossible to investigate all possible initial subsets, $\binom{n}{n_g}$ in number, even for small $n$ and $n_g$. Usually, we investigate a smaller number of say $J$ subsets of size $n_g$ each. A factor model is fitted to each subset giving a vector of estimated parameters $\hat{\boldsymbol{\theta}}_h = (\hat{\boldsymbol{\Lambda}}_h, \hat{\boldsymbol{\Psi}}_h, \hat{\boldsymbol{\Phi}}_h), h = 1, \ldots, J$.

Criteria for choosing the initial "basic" set among the $J$ subsets must be established. Those criteria are called objective functions, $F(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_h)$, and they are computed for each subset. The subset that yields the "optimum" value for the objection function is then selected. Different objective functions have been used in the literature, such as the minimum volume ellipsoid (MVE;

Hadi, 1992) and Mahalanobis distances (Atkinson et al., 2004). Both methods examine characteristics of subjects without taking into account the model. Here we propose to use measures that take the model parameters into account. More specifically, Rousseeuw (1984) estimated a regression model by minimizing the median of squares of the residuals using an algorithm that investigates many subsets of the data. According to his method, the objective function takes into account all observations, but model parameters are estimated from a subset of the data. Atkinson and Riani (2001) use a similar approach for extracting the initial subset for generalized models. Similarly here, we use a procedure for choosing the initial subset that resembles the Least Median of Squares Regression.

We propose to use an objective function that is based on the likelihood contributions of each observation. More specifically, we estimate the parameters of the assumed model in each subset $h$ ($h = 1, \cdots, J$). For each set of parameter estimates, we compute the median of the absolute likelihood contributions for the whole data set. Finally, we select the subset that has the minimum median of likelihood contributions. This criterion is denoted by *medlc*.

Alternatively, one can select the subset with the maximum log-likelihood value. That criterion is denoted with *maxl*. We have also used here as an objective function the minimum ULS criterion defined in Equation (5) that is denoted by *uls*.

Atkinson (1994) suggested repeating the FS starting from a number of different randomly chosen initial subsets. That will help checking the consistency of the method and identify outliers even if those were included in the initial subset.

## Progressing in the Forward Search

After fixing the size of the initial outlier-free subset ("basic" set) to $n_g$, there is a maximum of $n - n_g$ remaining steps in the FS algorithm up to which all observations will be included in the "basic" set. Note that at least one observation will be added in the "basic" set at each iteration.

At iteration $l$, a $q$-factor model is fitted to the "basic" set ($\boldsymbol{Y}_{l-1}^{bs}$). The *standard* FS orders all $n$ observations according to their closeness to the "basic" set. Closeness is based on criteria that use the model estimates from $\boldsymbol{Y}_{l-1}^{bs}$. Finally, the $(n_g + l)$ observations closest to the "basic" set are selected. This allows observations to enter and leave the "basic" set at each step of the FS. Alternatively, one can order only the observations in the "non-basic" set ($\boldsymbol{Y}_{l-1}^{nbs}$) by their "closeness" to the "basic" set ($\boldsymbol{Y}_{l-1}^{bs}$). The *standard* FS is practically more difficult to monitor because observations can enter and leave the "basic" set at any step. The two procedures were found through simulations to produce similar results when a large number of initial subsets is explored.

In regression applications, Hadi and Simonoff (1993) and Atkinson (1994) use studentized residuals as criteria for sorting the observations whereas Atkinson and Riani (2000) use raw residuals. We could also order observations in the "non-basic" set according to their Mahalanobis distances (MD) from the "basic" set where location and dispersion parameters are estimated from the "basic" set. Poon and Wong (2004) proposed a criterion that measures the effect of each observation from the "non-basic" set on the sample covariance matrix estimated in the "basic" set.

The criteria used here for progressing in the search are likelihood contributions, residuals, and MDs. Therefore, the subject with the largest likelihood contribution, or the smallest residual or the smallest MD will be the one to move from the "non-basic" set ($\boldsymbol{Y}_{l-1}^{nbs}$) to the "basic" set ($\boldsymbol{Y}_{l-1}^{bs}$). That subject is supposed to be more likely to be generated by the $q$-factor model fitted to the "basic" set.

## Monitoring the Search

The last step of the FS involves monitoring statistics of interest throughout the process using forward plots. By adding one observation at a time to the "basic" set one can monitor at each step of the search the effect that the addition of an observation has on parameter estimates, t-statistics, residuals, and goodness-of-fit tests and measures. More specifically, monitoring helps us identify observations that are not fitted by the hypothesized model.

*Residuals.* Residuals are frequently monitored during the FS (Atkinson, 1994; Atkinson & Riani, 2000, 2001). Jöreskog (1962) and Bollen and Arminger (1991) discuss residuals for linear factor models. There are two types of residuals that one may want to compute, namely, subject and aggregate residuals. Subject residuals measure deviation between an observed and an expected response under the fitted model to a particular variable or a set of variables. Aggregate residuals are defined as the difference between the fitted and (sample) observed covariance/correlation matrix. Because in this article we are interested in detecting atypical response patterns and not poorly fitted covariances we focus on subject residuals.

The residual for the $m^{th}$ individual and the $i^{th}$ manifest variable is given by

$$e_{mi} = y_{mi} - \sum_{j=1}^{q} \hat{\lambda}_{ij} z_{mj}, \quad m = 1 \ldots n, \quad i = 1, \ldots, p, \tag{2}$$

or in vector notation $\boldsymbol{e}_m = \boldsymbol{y}_m - \hat{\boldsymbol{\Lambda}} \mathbf{z}_m$, where $\boldsymbol{e}_m$ is of dimension $p \times 1$.

Note that because the latent variables are unobserved, estimated values $\hat{z}_{mj}$ are used instead of $z_{mj}$ in Equation (2). The $\boldsymbol{z}$ values can be replaced with factor scores (McDonald & Burr, 1967). The Bayesian Expected a Posteriori (EAP) score defined as the posterior mean of the latent variable ($E(z_j \mid \boldsymbol{y}_m)$) (Bartholomew, 1981) is used here and those residuals will be called in the article as EAP residuals. The residuals used here are a special case of the residuals discussed in Bollen and Arminger (1991).

A summarized across variables residual measure for an individual $m$ is given by

$$\boldsymbol{e}_m{}'\hat{\boldsymbol{\Psi}}^{-1}\boldsymbol{e}_m. \tag{3}$$

Bollen and Arminger (1991) refer to this type of standardization as a naive standardization because the variance of the predicted response is not taken into account. Because for the FS we are not interested in the distribution of the residuals, we use the naive ones that are easy to compute.

*Goodness-of-fit statistics and measures of fit.* The log-likelihood ratio test statistic (LR-statistic) is defined as

$$LR-statistic = n\left\{trace\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S} - \log|\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}| - p\right\}, \tag{4}$$

where $S$ and $\hat{\boldsymbol{\Sigma}}$ are the sample and estimated covariance matrices, respectively. Under certain conditions, the LR-statistic has a $\chi^2$ distribution. In the FS, the model is fitted successively to truncated samples and therefore the asymptotic distribution of the LR-statistic does not hold. We use here the LR-statistic as a goodness-of-fit measure rather than as a test statistic.

We also monitor the ULS statistic and the Root-Mean error Residual (RMR). The ULS is defined as

$$\text{ULS} = \frac{1}{2}\text{tr}(\mathbf{S}-\boldsymbol{\Sigma})^2, \tag{5}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'+\boldsymbol{\Psi}$ and $\mathbf{S}$ are the theoretical and sample covariance matrices, respectively. The RMR is defined as

$$RMR = \left[2\sum_{i}^{p}\sum_{j}^{i}(s_{ij}-\hat{\sigma}_{ij})^2/(p(p+1))\right]^{1/2}, \tag{6}$$

where $s_{ij}$ and $\hat{\sigma}_{ij}$ are elements of the sample and estimated under the factor model covariance matrices.

*Parameter estimates.* The effect of outliers on the parameter estimates can be checked by monitoring in the FS changes in the estimated factor loadings $\hat{\lambda}_{ij}$ as well as in the estimated variances of the unique factors ($\hat{\Psi}_i$). Alternatively, one can compute an overall measure of change in the matrix $\hat{\mathbf{\Lambda}}$. This is based on Cook's statistic, which is derived from the confidence region of the vector of all model parameters (Cook & Weisberg, 1982). Atkinson and Riani (2000) proposed a "forward version" of Cook's statistic,

$$D_m = (\hat{\boldsymbol{\lambda}}_{m-1} - \hat{\boldsymbol{\lambda}}_m)'\{cov(\hat{\boldsymbol{\lambda}}_{m-1})\}^{-1}(\hat{\boldsymbol{\lambda}}_{m-1} - \hat{\boldsymbol{\lambda}}_m), \tag{7}$$

where $\hat{\boldsymbol{\lambda}}_{m-1}$ is a vector obtained by stacking row by row the elements of the estimated loading matrix $\hat{\mathbf{\Lambda}}$ obtained from the "basic" set before the inclusion of the $m^{th}$ observation, $\hat{\boldsymbol{\lambda}}_m$ is the same vector but with observation $m$ now being included in the "basic" set, and $cov(\hat{\boldsymbol{\lambda}}_{m-1})$ is the estimated covariance matrix of the factor loadings obtained from an approximation of the inverse of the information matrix evaluated at the maximum likelihood solution.

*Outlier detection errors.* There are two kinds of errors that may occur in the process of detecting outliers, namely, the *masking* and the *swamping* effect. The *masking* effect occurs when an outlier is undetected because of the presence of a cluster of outliers and the *swamping* effect occurs when a "good" observation is incorrectly identified as an outlier. Under the masking effect, the importance of the observations is not evident unless several observations are deleted at once (Atkinson, 1986). Furthermore, a cluster of outliers will shift the mean from its true value and there is the possibility that some "good" observations may be classified as outliers (*swamping* effect).

## EXAMPLES

The FS will be applied to three data sets. As we have already discussed, different criteria are available for selecting the initial subset and for progressing in the FS. Though various criteria have been tried with our examples, it is not feasible to report all of them here. Running many forward searches both from randomly chosen and outlier-free initial samples, we concluded that likelihood contributions are a stable and efficient criterion for progressing in the search. All criteria, both for the selection of the initial sample and for the addition of subjects during the search, gave similar results.

For some of the examples we have also conducted a backward search. The backward search starts by fitting the model to the whole data and proceeds by deleting one observation at a time. The backward procedure is more likely to fail in the presence of "masked" outliers.

## Example 1: Open-Book and Closed-Book Examination Results

The first example is the open/closed book data set from Mardia, Kent, and Bibby (1979). This example involves exam grades of $n = 88$ students in mechanics, vectors, algebra, analysis, and statistics. The first two exams were given with closed books and the last three were given with open books. Lee and Wang (1996) studied the sensitivity of the one-factor model to minor perturbations of the data. They found that the deletion of Cases 81, 87, and 88 have a large effect on the hypothesized model with Case 81 being the most influential. However, Yuan and Bentler (1998b) found that Case 81, though it is the most influential, is not really atypical. Tanaka and Odaka (1989) analyzed this data set by examining the influence function for the common variance matrix $\mathbf{\Lambda}\mathbf{\Lambda}'$ and the unique variance matrix $\mathbf{\Psi}$ under a one-factor model. They found that the most influential cases are 75 and 82.

We examined 10,000 initial subsets of size 16 ($n_g = 16$). The criterion used for the selection of the initial subset is $maxl$. In Figure 1, forward plots of the LR-statistic defined in Equation (4), its asymptotic $p$ value, the ULS criterion in

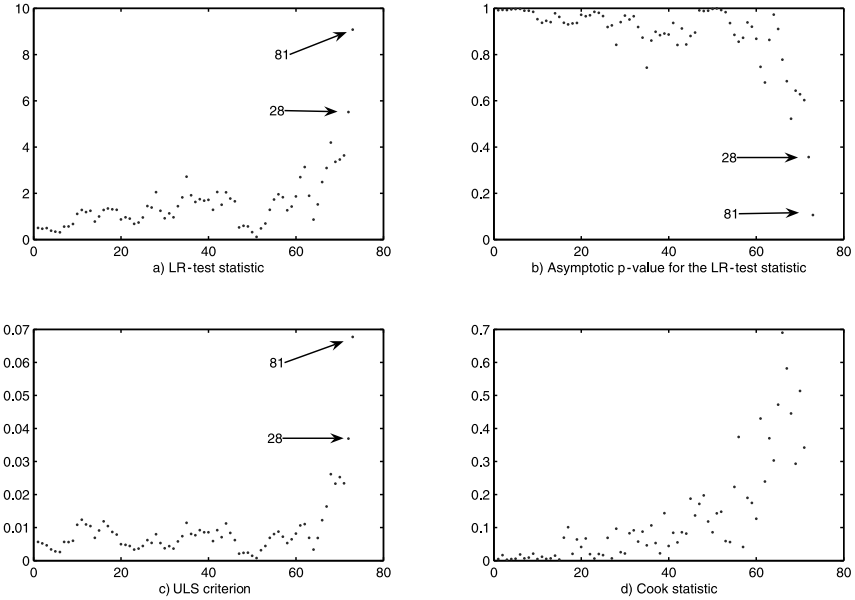


FIGURE 1 Example 1: Forward plots of the Log-Likelihood Ratio Statistic, its asymptotic $p$-value, the Unweighted Least Squares Statistic and the Root Mean Error Residual (the iterations of the forward search are shown on the horizontal axes and the statistic being monitored on the vertical axes).

Equation (5), and Cook's statistic defined in Equation (7) are given. It is obvious that the fit of the model deteriorates in the last steps of the search. Figure 1b shows clearly that the asymptotic $p$ value of the LR-statistic drops from 0.6 to 0.11 when Observations 28 and 81 enter the "basic" set. The exclusion of the last eight cases (33, 23, 66, 61, 54, 87, 28 and 81) give a $p$ value equal to 0.91. The low $p$ values are indicative of the large effect of those cases to the overall fit of the model. Figure 1d shows that cases that enter the FS toward the end of the search have a large effect on Cook's statistic and consequently on the parameter estimates.

We also conducted a backward search. The backward search starts with the whole sample and the observation with the lowest contribution to the log-likelihood is deleted. Parameters are reestimated after the deletion of an observation. Subjects were omitted in the following order: 81, 28, 87, 54, 61, 66, 23, 33. In that example, there is an agreement between forward and backward procedures.



## Example 2: An Artificial Data Set With Masked Multivariate Outliers

From the previous example, we saw that the FS provides a natural ordering of the data with outliers entering toward the end. In addition, both the backward and the forward search indicated the same outliers. We use a simulated data set to investigate the performance of the FS in the presence of masked multivariate observations. By masked outliers we mean that there is a cluster of outliers that dominates the analysis to that extent that the model fits that cluster satisfactorily. These outliers do not necessarily have large MDs and large residuals.

An artificially contaminated data set of five items and 100 individuals is created as follows:

- A sample of 80 individuals is generated from a multivariate normal distribution with zero mean vector, variances equal to unity, and correlations drawn at random from a uniform distribution within the interval [0.4,0.7].
- Twenty subjects are generated from a standard multivariate normal distribution (mean vector zero and identity variance-covariance matrix). Their absolute values were taken. We consider those 20 subjects contaminated data.
- Contaminants are labeled from 1 to 20.

A one-factor model is fitted and a backward and a forward search are conducted. From Figure 2, we see that none of the contaminated cases are identified as having large residuals. That indicates that the generated outliers are masked. Observations 39, 82, and 93 have the largest residuals.

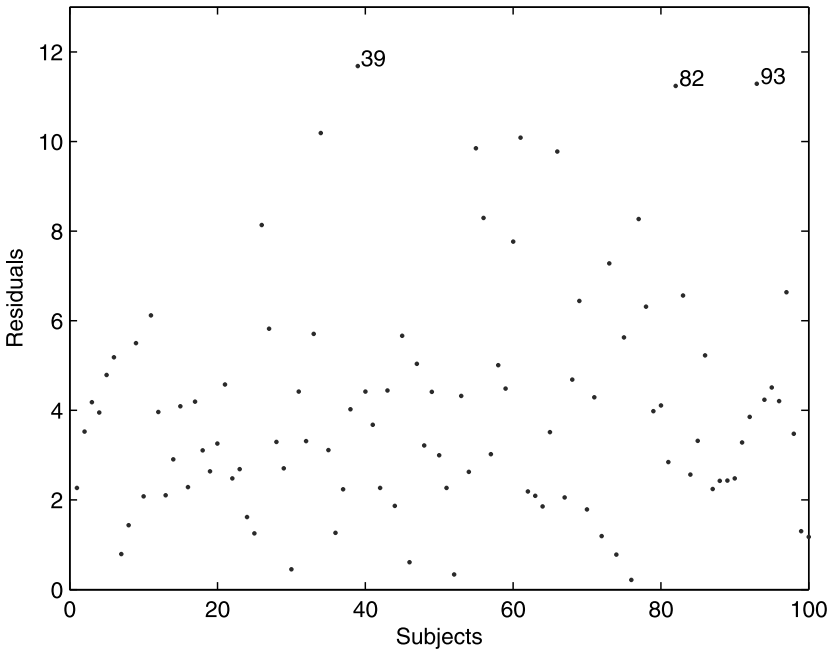


FIGURE 2 Example 2: Plot of Expected a Posteriori residuals, one-factor model.

Figure 3a plots the MDs whereas Figure 3b plots the robust MDs for which the MVE is used for the estimation of robust means (location parameter) and a robust covariance matrix (dispersion or scatter parameter). Rousseeuw and Van Zomeren (1990) suggested comparing the robust distances with the 0.975 quantile of a $\chi^2_p$. In our data set, $p = 5$. The cutoff value is plotted in both figures. None of the 20 contaminated cases were found to lie outside the cutoff value. From Figure 3a, Case 39 is shown as an outlier whereas from Figure 3b a plethora of outliers is shown. This is in agreement with the discussion of Cook and Hawkins that appears in Rousseeuw and Van Zomeren (1990) that the MVE yields a plethora of outliers.

We first conducted a backward search on the simulated data set. The first 20 individuals that are excluded from the "basic" set are 39, 93, 83, 82, 55, 66, 61, 56, 34, 78, 45, 26, 60, 75, 69, 68, 27, 49, 77, and 43. Note that none of the contaminated cases are excluded. This is an indication that the outliers are masked and not detected by the backward search.

We proceed by applying the FS to the simulated data set. The initial sample was selected using the ULS criterion after examining 1,000 subsets of size 16. The initial "basic" set did not contain any of the 20 contaminated cases. The

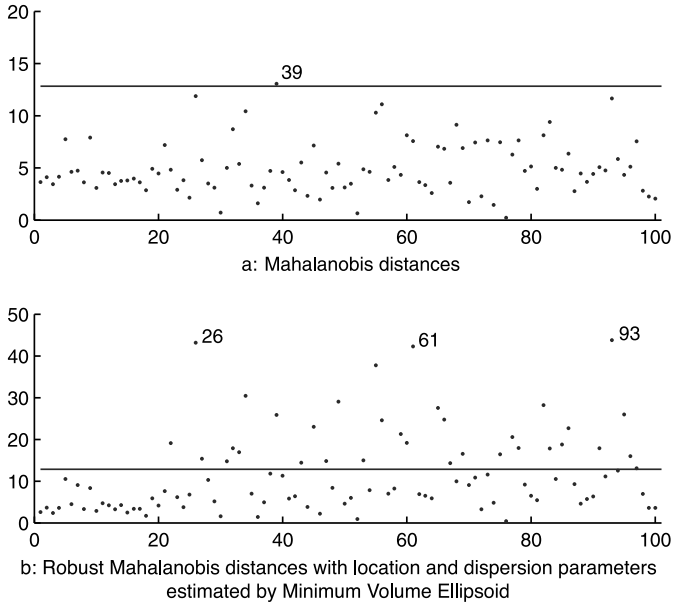


FIGURE 3 Example 2: Mahalanobis and robust Mahalanobis distances (the horizontal axes represent the observations and the vertical axes represent the Mahalanobis distances).

search consists of 84 $(100 - 16)$ steps. Progressing in the search was based on likelihood contributions.

The 20 contaminants started entering the "basic" set at Step 51 with the last entering at Step 79. Despite the fact that outliers did not enter at the end of the search, we can still detect them from the sharp changes of the statistics monitored in the forward plots. Figure 4 shows various forward plots that assess the fit of the model. Figures 4a and 4b give forward plots of goodness-of-fit measures. The LR-statistic and the ULS are computed from Equation (4) and Equation (5), respectively. A deterioration of the fit after Iteration 51 is obvious. Figure 4c gives the forward plot of Cook's statistic defined in Equation (7). It is evident that the inclusion of outliers substantially influences the parameter estimates. In Figure 4d, we see that two of the five estimated variances of the error terms $(\hat{\Psi}_2, \hat{\Psi}_3)$ increase dramatically and start approaching unity after the inclusion of the contaminants. The most interesting plots are 4e and 4f that refer to the EAP residuals. Figure 4e shows the forward plot of residuals for Subjects 39, 93, and 83. These are the first three observations to leave the "basic" set in a backward search and therefore they are potentially outliers. However, a forward plot of residuals indicates a sharp increase after Step 50. These observations are mistakenly considered outliers after Step 50 with the inclusion of the artificial

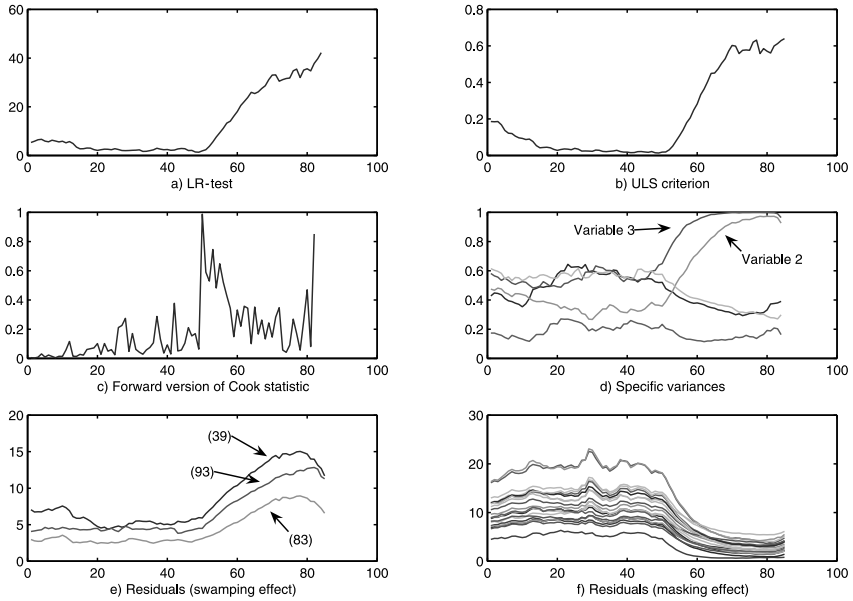


FIGURE 4 Example 2: Forward plots of the Log-Likelihood Ratio Statistic, the Unweighted Least Squares Statistic, the Cook Statistic, the specific variances and the Expected a Posteriori Residuals, one-factor model (the iterations of the forward search are shown on the horizontal axes and the statistic being monitored on the vertical axes).

contaminants. This is an example of the *swamping* effect. Figure 4f refers to the EAP residuals for the contaminated cases that maintain large values until their inclusion, and from that point on they become masked. Such plots are typical in forward searches and analogous residual plots can be found in regression analysis (Atkinson & Riani, 2000).

## Example 3: Car Example

With the third example, we investigate the performance of the FS when the wrong model is specified. Riani and Atkinson (2000) use the FS for selecting the right transformation of the dependent variable in regression analysis models and Atkinson and Riani (2001) use the FS for determining the appropriate link function in generalized linear models. In both cases, FS is used not only as an outlier detection technique but also as a model selection technique.

So far we applied the FS to data sets that were poorly fitted by a one-factor model. In Example 1, the fit of the one-factor model was considerably improved by omitting a few cases. In some examples, only a fraction of cases may be

responsible for the introduction of a second factor. The matter of interest is how the FS would behave when the entire data set, and not just a fraction of it, is generated by a more complex model than the one assumed when applying the FS. To explore that situation, we took a data set that was satisfactorily fitted by a two-factor model and ran an FS under the one-factor model.

We used a data set consisting of five traits (items) on 391 cars. The items are Acceleration, Displacement, Horsepower, Miles Per Gallon, and Weight. The data set can be found at http://lib.stat.cmu.edu/datasets/cars.data

The LR-statistic of the one-factor model is 259.748 ($p$ value $< .000$). A two-factor model gives an LR-statistic of 1.137 ($p$ value $= .286$), showing a significant improvement over the one-factor model.

We conducted an FS when the one-factor model was fitted. The size of the initial subset was chosen to be $n_g = 50$ and the criterion used for selecting the initial subset among 1,000 different subsets was the one that yields the maximum likelihood (*maxl* criterion). Figure 5 gives a forward plot of the LR-statistic (dotted line) together with 95% simulation envelopes (solid lines) produced from 100 repetitions of the search and when the one-factor model was fitted. Even

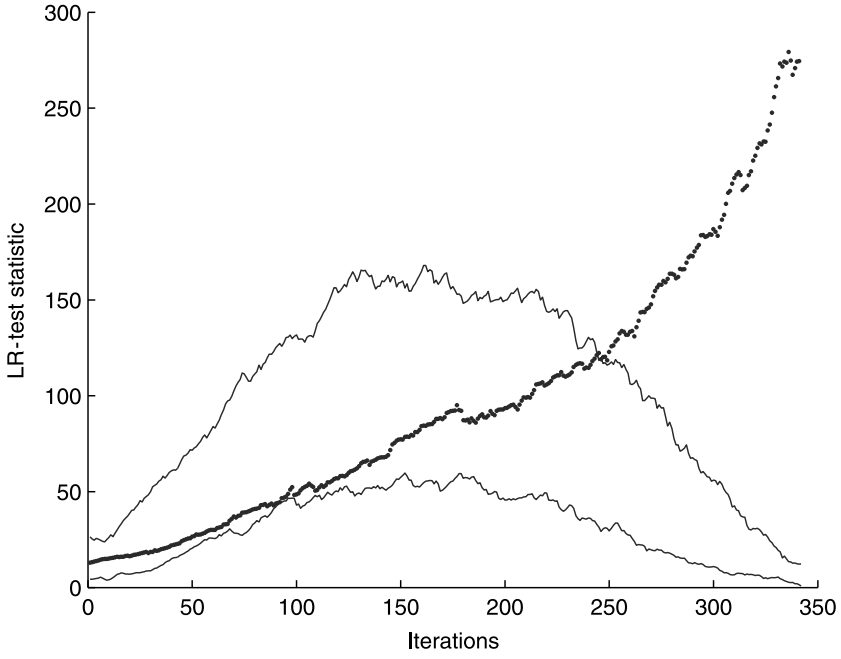


FIGURE 5 Example 3: Forward plot of the Log-Likelihood Ratio Statistic with 95% simulation envelopes ($n_g = 50$, one-factor model).

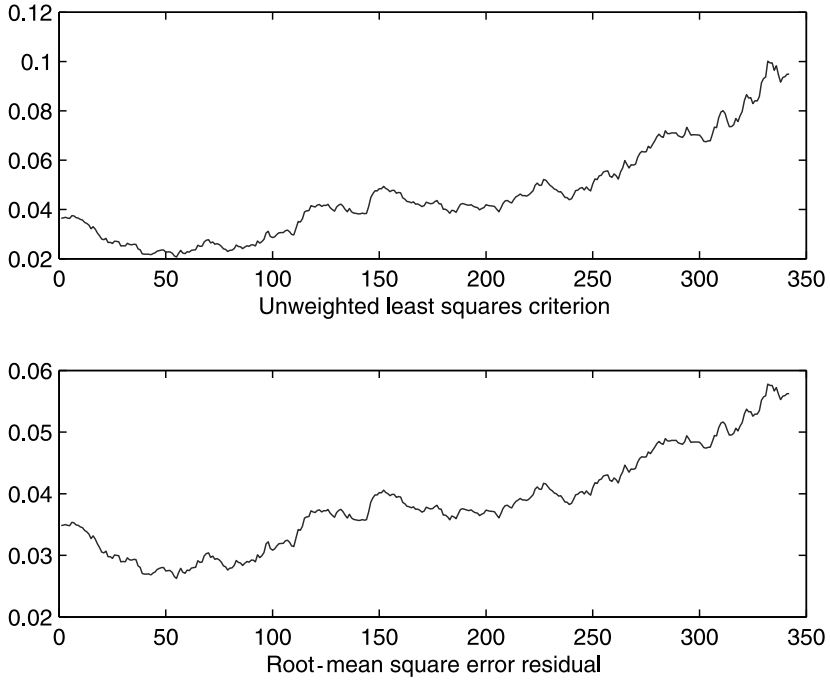


FIGURE 6 Example 3: Forward plot of Unweighted Least Squares and Root Mean Error Residual criteria ($n_g = 50$, one-factor model, horizontal axes represent the iterations of the Forward Search and vertical axes represent the statistic being monitored).

in the beginning of the search subsets are poorly fitted by the one-factor model and the fit deteriorates even more with the progress of the search. In the last 100 iterations, the values of the LR-statistic lie outside the upper 97.5% line of the simulation envelope, indicating a poor fit. The ULS criterion and the RMR are monitored in Figure 6. Both indices reveal a deteriorating fit as observations are added to the "basic" set. Figure 7 shows that throughout the search, parameter estimates did not vary much. Finally, a forward plot of the LR-statistic under a two-factor model is given in Figure 8. The LR-statistic increases toward the last steps of the FS.

## SIMULATION STUDY

In this section, we conduct simulation studies that aim to evaluate the criteria used in the different steps of the FS. Specifically, we focus on the power of various criteria used for extracting outlier-free initial subsets and on the ability

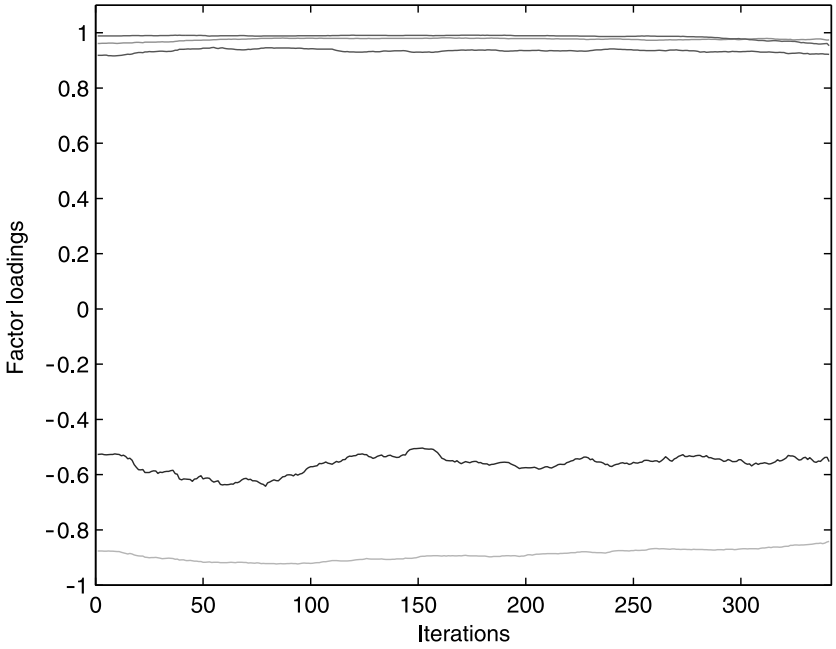


FIGURE 7 Example 3: Forward plot of the estimated factor loadings (initial sample size = 50, one-factor model).

of FS to detect outliers. Atypical response patterns are detected either when they enter in the last steps of the search or when sharp changes are shown in the statistics being monitored.

Our first goal is to evaluate the criteria (objective functions) used for the extraction of an outlier-free initial subset. Data are generated from a one-factor model and a small proportion of observations is contaminated. Four different contamination schemes have been investigated. The first two contamination schemes are merely point mass contamination where arbitrary large values are set to a proportion of the data, the third scheme is typical model deviation contamination where a small proportion of the data is generated from a model with different mean values from the hypothesized one, and in the fourth contamination scheme a proportion of uncorrelated observations with the same means and variances as our data has contaminated the data. In the fourth contamination scheme only the correlation structure of the contaminants differs substantially from the correlation structure of the data.

The contaminants in the first three schemes are set symmetrically to the mean. It is not typical to observe subjects with large values on some of the items and low values on some other items while the items are positively correlated.

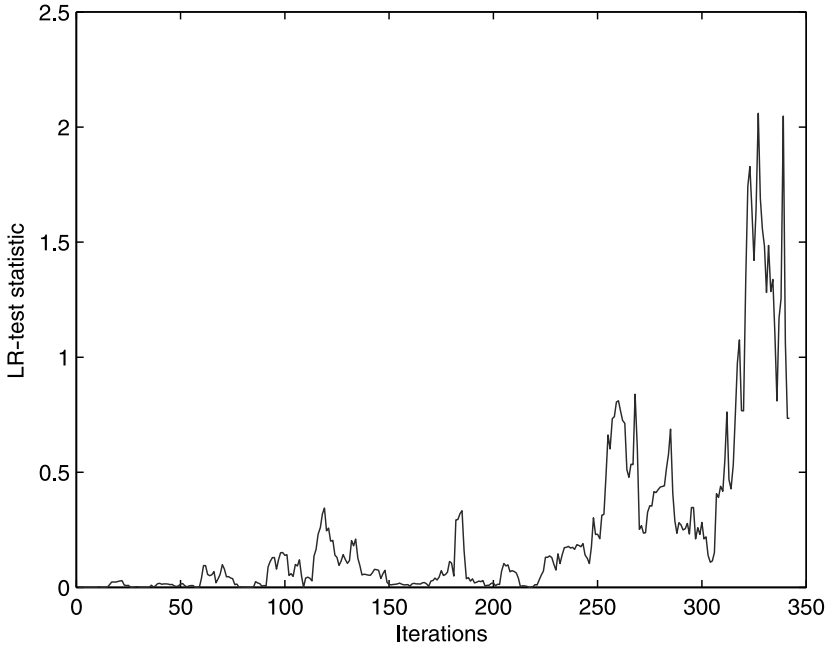


FIGURE 8 Example 3: Forward plot of the Log-Likelihood Ratio Statistic under a two-factor model (initial sample size = 50, two-factor model).

There are innumerable combinations of sample size and number of manifest items as well as model parameters. We choose the following simulation scheme with $n = 120$ and $p = 5$.

The four contamination schemes, after variables have been standardized, can be summarized as follows:

1. A small proportion of the data and only two variables are set to arbitrary large values. The contaminated values chosen for the two variables are the same but with opposite sign. That contamination will be denoted by $C_1(o)$ where $o$ is the arbitrary value used.
2. Arbitrary large values are set to a small proportion of the data and to all five variables. One of the variables has the same arbitrary value but with opposite sign. This contamination will be denoted by $C_2(o)$.
3. The values of four variables for a small proportion of the data are generated from a normal distribution with mean equal to $o$ and unit variance. The values of the fifth variable are generated from a normal distribution with mean value $-o$. This form of contamination will be denoted by $C_3(o)$

where $o$ is the mean value of the normal distribution that generated the contaminants.

4. A small proportion of the data is generated from a multivariate normal distribution with the same mean vector as the rest of the data and the identity matrix as variance-covariance matrix. This way of contamination will be denoted by $C_4$.

A number $J$ of subsets each of size $n_g$ is investigated for the extraction of the initial subset. Various objective functions for the selection of the initial subset have been discussed in the article such as the $medlc$ that selects the initial subset with the minimum of the median of likelihood contributions, the $maxl$ that selects the initial subset with the maximum log-likelihood value, and the ULS that selects the subset with the lowest value for the ULS statistic. In Tables 1 to 3, the column labeled "Power" gives the percentage of times an outlier-free initial subset is extracted. The first column gives the contamination scheme and the second column gives the percentage of contamination. The column labeled "expected" gives the expected number of outlier-free subsets under each simulation scheme when the initial subset is chosen randomly (without using an objective function). The expected probability is computed by the hypergeometric distribution and then it is multiplied by $J$ to get the expected number of clean subsets.

Table 1 shows that the power of the objective function $medlc$ tends to be higher under all four contamination schemes when the number of initial subsets

TABLE 1
Power of the objective function *medlc*, $n = 120$, $p = 5$

| *Contamination Scheme* | *Proportion of Contamination* | $J$ | $n_g$ | *Expected* | *Power* |
|---|---|---|---|---|---|
| $C_1(1.5)$ | 5 | 100 | 30 | 17 | 84 |
| $C_1(1.5)$ | 10 | 100 | 30 | 2.6 | 43 |
| $C_1(1.5)$ | 10 | 100 | 16 | 16.4 | 83 |
| $C_1(1.5)$ | 10 | 1000 | 30 | 26 | 72 |
| $C_1(1.5)$ | 10 | 1000 | 16 | 164 | 100 |
| $C_2(2)$ | 10 | 100 | 30 | 2.6 | 89 |
| $C_2(2)$ | 10 | 1000 | 30 | 26 | 95 |
| $C_2(3)$ | 10 | 100 | 30 | 2.6 | 96 |
| $C_2(3)$ | 10 | 1000 | 30 | 26 | 100 |
| $C_3(1.5)$ | 10 | 100 | 30 | 2.6 | 86 |
| $C_3(3)$ | 10 | 100 | 30 | 2.6 | 99 |
| $C_4$ | 10 | 100 | 30 | 2.6 | 37 |
| $C_4$ | 10 | 1000 | 30 | 26 | 81 |

TABLE 2
Power of the objective function *maxl*, $n = 120$, $p = 5$

| *Contamination Scheme* | *Proportion of Contamination* | $J$ | $n_g$ | *Expected* | *Power* |
|---|---|---|---|---|---|
| $C_1(1.5)$ | 10 | 100 | 30 | 2.6 | 67 |
| $C_1(1.5)$ | 10 | 1000 | 30 | 26 | 87 |
| $C_1(1.5)$ | 10 | 100 | 16 | 16.4 | 85 |
| $C_2(1.5)$ | 10 | 100 | 30 | 2.6 | 71 |
| $C_2(1.5)$ | 10 | 1000 | 30 | 26 | 98 |
| $C_3(1.5)$ | 10 | 100 | 30 | 2.6 | 100 |
| $C_4$ | 10 | 100 | 30 | 2.6 | 86 |
| $C_4$ | 20 | 1000 | 16 | 26 | 100 |

tested increases and when the size of the initial subset is reduced. Tables 2 and 3 show similar results for the objective functions $maxl$ and ULS, respectively. Note that in practice, it is very time-consuming to investigate a large number of subsets and also the model estimation becomes unstable for small sample sizes.

We also attempt to assess the ability of the FS in detecting contaminants with a small simulation. An FS with $medlc$ used as the objective function for extracting the initial subset among a set of 100 subsets and the likelihood contributions used as a method of progressing in the search are conducted for 10 data sets of size $n = 100$ and $p = 5$. Each data set is generated from a one-factor model. Six percent of the data has been contaminated under scheme $C_2(2)$. In Figure 9 we monitor the largest EAP residual from the "basic" set at each step of those 10 forward searches. In all searches the six contaminants enter in the last six iterations. It is evident from Figure 9 that the largest EAP residual

TABLE 3
Power of the objective function *uls*, $n = 120$, $p = 5$

| *Contamination Scheme* | *Proportion of Contamination* | $J$ | $n_g$ | *Expected* | *Power* |
|---|---|---|---|---|---|
| $C_1(1.5)$ | 5 | 100 | 30 | 17 | 94 |
| $C_2(1.5)$ | 5 | 100 | 30 | 17 | 99 |
| $C_2(1.5)$ | 10 | 100 | 30 | 2.6 | 71 |
| $C_2(1.5)$ | 10 | 1000 | 30 | 26 | 99 |
| $C_2(1.5)$ | 10 | 100 | 16 | 16.4 | 98 |
| $C_4$ | 10 | 1000 | 16 | 164 | 98 |
| $C_4$ | 20 | 1000 | 16 | 21 | 89 |

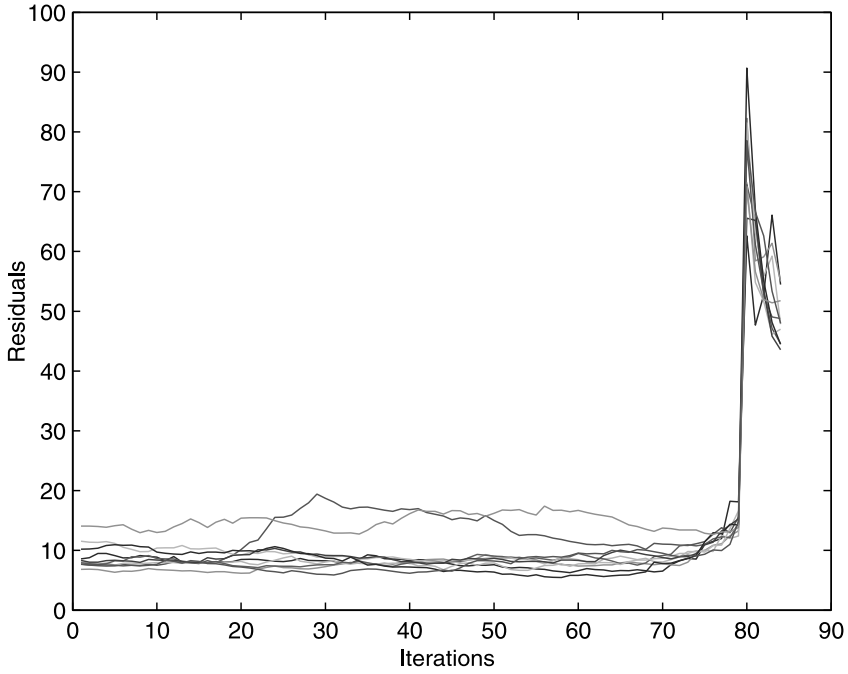


FIGURE 9 Largest Expected a Posteriori residuals in 10 forward searches.

in all 10 searches increased drastically after the inclusion of the contaminants and then started decreasing again after the inclusion of the first few outliers. This is attributed to masking. Similar results are reached when we consider various contamination schemes with different criteria used for the extraction of the initial subset and for progressing in the search.

## CONCLUSIONS

FS provides an easy and robust way of detecting contaminants (outliers, aberrant response patterns) in factor analysis. Unlike robust factor analysis, the robustness of the FS algorithm is not achieved by finding an estimator with bounded influence function and a large breakdown point. The breakdown point of the FS estimator is unknown but we agree with Atkinson and Riani (2000) that this in not a drawback to the procedure. Contaminants are easily detected by sharp changes in forward plots of functions of model parameters. Forward plots

provide valuable information regarding the structure of the data and deviations from the hypothesized model.

Outliers tend to enter the "basic" set in the last iterations of the search when there are no masked outliers and the initial subset is outlier free. Even when the initial subset contains outliers, forward plots of residuals and goodness-of-fit measures will show sharp changes that would make us suspect the existence of contaminants in the initial sample. We have also shown that FS has an advantage over backward methods in the presence of a cluster of outliers.

We have also conducted a *standard* FS in all our examples. The *standard* FS orders all $n$ subjects at each step and not just the ones from the "non-basic" set. The advantage of the *standard* FS is that makes the selection of an outlier-free initial subset less important but at the same time makes it more difficult to monitor the statistics of interest, especially when the interchange of observations between the "basic" and "non-basic" set is severe. In all our examples, we did not observe severe interchange among observations between the "basic" and the "non-basic" set when a *standard* FS was applied.

Various statistics can be used for selecting the initial subset and progressing in the search. The various combinations checked led to the conclusion that all methods are powerful in extracting clean initial samples and the differences in the ordering of data when different progressive criteria have been used are negligible. All methods provide a natural ordering of the data and smooth forward plots. However, the reliability of most measures being monitored is unknown. There are various types of residuals and goodness-of-fit statistics that need further evaluation.

In the early steps of the search, where the sample size is small, it is likely to get improper solutions (negative error variances). When improper solutions occur one needs to run more forward searches.

The behavior of goodness-of-fit statistics under improper solutions is unknown. Some modifications of the FS when it is applied to large data sets are necessary. When the sample size and the number of variables is large efficient algorithms need to be established for selecting the initial subset. Repeating the FS by randomly selected initial subsets (Atkinson, 1994) is helpful not only for locating the contaminants but also for studying the consistency of the FS and its behavior when contaminants are present in the initial subset.

The FS can be applied to any latent variable model. The purpose of the algorithm is to identify observations that have not been generated by the hypothesized model. Therefore, the type of model used does not change the steps and the rationale of the criteria that should be used to progress and monitor the search. As a result the algorithm can be applied to either exploratory or confirmatory factor analysis models and to models with categorical responses. Appropriate criteria (objective functions) and statistics to be monitored need to be developed for the categorical case.

## REFERENCES


Atkinson, A. C. (1986). Masking unmasked. *Biometrika, 73*, 533–541.

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association, 89*, 1329–1339.

Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York: Springer.

Atkinson, A. C., & Riani, M. (2001). Regression diagnostics for binomial data from the forward search. *The Statistician, 50*, 63–78.

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer-Verlag.

Bartholomew, D. J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology, 434*, 93–99.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology, 21*, 235–262.

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics, 29*(3), 231–237.

Cheng, T.-C., & Victoria-Feser, M.-P. (2002). High-breakdown estimation of multivariate mean and covariance with missing observations. *British Journal of Mathematical and Statistical Psychology, 55*, 317–335.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman and Hall.

Davies, P. L. (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *Annals of Statistics, 15*, 1269–1292.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association, 76*(374), 354–362.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators* (Qualifying paper). Harvard University, Boston, MA.

Filzmoser, P. (1999). Robust principal component and factor analysis in the geostatistical treatment of environmental data. *Environmetrics, 10*, 363–375.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B, 54*(3), 761–771.

Hadi, A. S., & Simonoff, J. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association, 88*(424), 1264–1272.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics, the approach based on influence functions*. New York: Wiley.

Huber, P. J. (1981) *Robust statistics*. New York: Wiley.

Jöreskog, K. G. (1962). On the statistical treatment of residuals in factor analysis. *Psychometrika, 27*, 335–353.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–482.

Jöreskog, K. G. (1972). Factor analysis by generalized least squares. *Psychometrika, 37,* 243–260.

Jöreskog, K. G. (1979). Structural equation models in the social sciences: Specification, estimation and testing. In K. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 105–127). Cambridge, MA: Abt Books.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworth.

Lee, S.-Y., & Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika, 61,* 93–108.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics, 4,* 51–67.

McDonald, R. P., & Burr, E. J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika, 32,* 381–401.

Moustaki, I., & Victoria-Feser, M.-P. (2006). Bounded-influence robust estimation in generalized linear latent variable models. *Journal of the American Statistical Association, 101*(474), 644–653.

Pison, G., Rousseeuw, P. J., Filzmoser, P., & Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis, 84*(7), 145–172.

Poon, W.-Y., & Wong, Y.-K. (2004). A forward search procedure for identifying influential observations in the estimation of a covariance matrix. *Structural Equation Modeling, 11,* 357–374.

Riani, M., & Atkinson, A. C. (2000). Robust diagnostic analysis: Transformations in regression. *Technometrics, 42,* 384–394.

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 W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and applications* (pp. 283–297). Dordrecht, The Netherlands: Reidel.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*, 633–651.

Stahel, W. A. (1981). *Robust estimation: Infinitesimal optimality and covariance matrix estimators*. Unpublished doctoral dissertation, ETH, Zurich, Switzerland.

Tanaka, J. S., & Odaka, Y. (1989). Influential observations in principal factor analysis. *Psychometrika, 54*, 475–485.

Woodruff, D., & Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association, 89*, 888–896.

Yuan, K. M., & Bentler, P. M. (1998a). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 51*, 63–88.

Yuan, K. M., & Bentler, P. M. (1998b). Structural equation modeling with robust covariances. *Sociological Methodology, 28*, 363–396.

Yuan, K. M., & Bentler, P. M. (2001). Effects of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 54*, 161–175.