Robust Estimation of Constrained Covariance Matrices for Confirmatory Factor Analysis

DUPUIS LOZERON, Elise, VICTORIA-FESER, Maria-Pia

Abstract

Confirmatory factor analysis (CFA) is a data analysis procedure that is widely used in social and behavioral sciences in general and other applied sciences that deal with large quantities of data (variables). The classical estimator (and inference) procedures are based either on the maximum likelihood (ML) or generalized least squares (GLS) approaches which are known to be non robust to departures from the multivariate normal assumption underlying CFA. A natural robust estimator is obtained by first estimating the (mean and) covariance matrix of the manifest variables and then "plug-in" this statistic into the ML or GLS estimating equations. This two-stage method however doesn't fully take into account the covariance structure implied by the CFA model. An S -estimator for the parameters of the CFA model that is computed directly from the data is proposed instead and the corresponding estimating equations and an iterative procedure derived. It is also shown that the two estimators have different asymptotic properties. A simulation study compares the finite sample properties of both estimators showing that the proposed [...]
Robust Estimation of Constrained Covariance Matrices for Confirmatory Factor Analysis

E. Dupuis Lozeron\textsuperscript{a}, M.P. Victoria-Feser\textsuperscript{a}

\textsuperscript{a}HEC, University of Geneva, 1211 Geneva, Switzerland

Abstract

Confirmatory factor analysis (CFA) is a data analysis procedure that is widely used in social and behavioral sciences in general and other applied sciences that deal with large quantities of data (variables). The classical estimator (and inference) procedures are based either on the maximum likelihood (ML) or generalized least squares (GLS) approaches which are known to be non robust to departures from the multivariate normal assumption underlying CFA. A natural robust estimator is obtained by first estimating the (mean and) covariance matrix of the manifest variables and then “plug-in” this statistic into the ML or GLS estimating equations. This two-stage method however doesn’t fully take into account the covariance structure implied by the CFA model. An $S$-estimator for the parameters of the CFA model that is computed directly from the data is proposed instead and the corresponding estimating equations and an iterative procedure derived. It is also shown that the two estimators have different asymptotic properties. A simulation study compares the finite sample properties of both estimators showing that the proposed direct estimator is more stable (smaller MSE) than the two-stage estimator.

Key words: $S$-estimator, asymptotic efficiency, generalized least squares, structural equations modeling, two-stage estimation

1. Introduction

Factor analysis (FA) is an old technique that is widely used not only as a data reduction technique, but also for the construction of measurement scales as is done for example in psychology. It is also used for the analysis of data in social and behavioral sciences in general and other applied sciences that deal with large quantities of data (variables). Given a vector of observed (manifest) variables $X = (X_1, ..., X_p)^T \in \mathbb{R}^p$ with mean vector $\mu = (\mu_1, ..., \mu_p)^T$ and covariance matrix $\Sigma$, the FA model states that

$$X - \mu = \Gamma f + \epsilon,$$

where $f = (f_1, ..., f_k)^T$ is the vector containing $k < p$ factors, $\Gamma$ is the $p \times k$ matrix of factor loadings and $\epsilon$ is the error term. It is also assumed that $\text{cov}(f, \epsilon) = 0$, $E[f] = 0_k$, $E[\epsilon] = 0_p$, $E[ff^T] = \Phi_{kk}$, $E[\epsilon\epsilon^T] = \Psi_{pp}$ and $\Psi = \text{diag}(\psi_1^2, ..., \psi_p^2)$, where the $\psi_j^2$ are the so-called

\textit{Preprint submitted to Computational Statistics and Data Analysis, August 20, 2009}
uniqueness. Moreover, since $E[(X - \mu)(X - \mu)^T] = E[(\Gamma f + \epsilon)(\Gamma f + \epsilon)^T]$, the FA model is also often presented as

$$\Sigma = \Gamma \Phi \Gamma^T + \Psi.$$  

(1)

The loadings matrix $\Gamma$, containing say $q$ unknown factor loadings $\gamma_1, \ldots, \gamma_q$, can be either full for so-called exploratory FA (i.e. $q = p \cdot k$) or constrained (typically some loadings are set to 0) for so-called confirmatory FA (i.e. $q < p \cdot k$). In what follows, we will assume that the factors are normal and independent, i.e. $\Phi = \mathbf{I}_p$, but this hypothesis could be relaxed.

Given a sample of $n$ observations on the $p$ manifest variables, with classical FA, $\mu$ is estimated by the sample mean $\bar{x}$ and $\Sigma$ is estimated by the sample covariance matrix $S$. Estimates for $\Psi$ and $\Gamma$ are then found by maximum likelihood estimator (MLE) algorithm as $S$ is supposed to have a Whishart distribution if $X$ is multivariate normal (see Jöreskog, 1967). However, it is well known that the sample mean and covariance are not robust estimators of their corresponding population parameters, in that small departures from the normality assumption on the manifest variables can seriously bias the FA parameter estimates (see e.g. Yuan and Bentler, 1998, 2001; Pison et al., 2003). This is also true for alternative estimators to the MLE, such as generalized least squares (GLS) methods (see Browne, 1984).

To avoid potential biases due to the violation of the normality assumption, one can simply estimate the mean and covariance of the manifest variables in a robust fashion and “plug in” the estimates in a classical objective function $L$ (to be minimized), such as the one for the MLE based on the Whishart distribution, i.e.,

$$L_{ML}(\theta) = \log|\Sigma(\theta)| + \text{tr}[\Sigma(\theta)^{-1}] - \log|S| - p,$$

(2)

or the one for the GLS

$$L_{GLS}(\theta) = (\hat{\sigma} - \sigma(\theta))^T W (\hat{\sigma} - \sigma(\theta)),$$

(3)

where $W$ is some weight matrix, $\theta$ is the vector of parameters of interest (the $\gamma_j$ and $\psi_j^2$), $\Sigma(\theta)$ depends on $\theta$ through (1) and $\sigma(\theta) = \text{vec}(\Sigma(\theta))$ and $\hat{\sigma} = \text{vec}(S)$, where the operator vec(.) stacks the columns of a matrix on top of each other. There are different possibilities for the choice of $W$, the one leading to the most efficient estimator of $\theta$ being the asymptotic covariance matrix of $\hat{\sigma}$. In practice, this quantity is replaced by a suitable estimate which possibly depends on the unknown parameters $\theta$. This type of procedure that has been formalized in e.g. Yuan and Bentler (1998) and Pison et al. (2003), provides consistent estimators of the FA model’s parameters. An alternative approach for robust estimation of FA models is also given in Croux et al. (2003).

What is less clear however is the efficiency property of the resulting estimators. Indeed, this two-stage method implies that a robust (mean and) covariance matrix is first estimated, without using the fact that under the FA model, the covariance matrix is constrained by (1). Yuan and Bentler (1998) propose to estimate the FA model’s parameters directly using a robust estimator but do not implement the method. In this paper, we also propose to estimate the parameter of the FA model directly in a robust fashion. For that, we use a constrained biweight $S$-estimator (CBS) in the same spirit as the constrained translated biweight $S$-estimator proposed by Copt and Victoria-Feser (2006) for mixed linear models. We develop an iterative algorithm to compute it in the framework of confirmatory FA. We also investigate its large and small sample properties in terms of variance (and bias) and compare its performance to a two-stage robust estimator using also the same $S$-estimator. We do this analytically and by means of simulations in different models and settings.
2. Robust estimators for FA

2.1. Two-stage estimator

To estimate \( \mu \) and \( \Sigma \), we use an \( S \)-estimator with a high breakdown point. Very generally, an \( S \)-estimator of multivariate location and scale is defined as the solution for \( \mu \) and \( \Sigma \) that minimizes the determinant of \( \Sigma \), i.e. \( |\Sigma| \), subject to

\[
\frac{1}{n} \sum_{i=1}^{n} \rho \left( \sqrt{(\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)} \right) = \frac{1}{b_0} \sum_{i=1}^{n} \rho(d_i),
\]

where \( \rho \) is a symmetric, continuously differentiable function, with \( \rho(0) = 0 \) and \( \rho \) is strictly increasing on \([0, c]\) and constant on \([c, \infty)\) and \( c \) is such that \( E_\Phi[\rho(c)] = \epsilon^* \) where \( \Phi \) is the standard normal and \( \epsilon^* \) is the chosen breakdown point (see Rousseeuw and Yohai, 1984). \( b_0 \) is a parameter chosen to determine the breakdown point by means of \( b_0 = \epsilon^* \max d \rho(d) \).

For \( \rho \), we use the biweight \( \rho \)-function of Tukey. Hence, the biweight \( S \)-estimator (BS) is defined by the minimum of \( |\Sigma| \) subject to (4), with

\[
\rho(d; c) = \begin{cases} 
\frac{d^2}{2} - \frac{d^4}{2(c^2)} + \frac{d^6}{6(c^4)}, & 0 \leq d \leq c \\
\frac{c^2}{6}, & d > c 
\end{cases}
\]

Estimating equations for the BS-estimator are

\[
\mu = \frac{1}{\sum_{i=1}^{n} u(d_i; c)} \sum_{i=1}^{n} u(d_i; c) \mathbf{x}_i,
\]

\[
\Sigma = \frac{p}{\sum_{i=1}^{n} u(d_i; c) d_i^2} \sum_{i=1}^{n} u(d_i; c) (\mathbf{x}_i - \mu) (\mathbf{x}_i - \mu)^T,
\]

with weights

\[
u(d; c) = \frac{1}{d} \frac{\partial}{\partial d} \rho(d; c) = \begin{cases} 
(1 - (d/c)^2)^2, & 0 \leq d \leq c \\
0, & d > c 
\end{cases}
\]

The BS-estimator of \( \Sigma \) can then be used to replace \( \mathbf{S} \) or \( \hat{\sigma} \) in the objective functions (2) or (3); this defines the two-stage estimator (MLE or GLS).

2.2. Direct estimator

The FA model implies that the data have been generated by a multivariate normal distribution with constrained covariance matrix according to (1). For this type of problems, one can use the results of Copt and Victoria-Feser (2006) who have developed such an estimator for mixed linear models.

In Appendix A, we derive estimating equations for the FA parameters directly from the objective function of the \( S \)-estimator, using the Lagrangian of the minimization of \( |\Sigma| \) subject to (4) given by

\[
\mathcal{L} = \log |\Sigma| + \lambda \left[ \frac{1}{n} \sum_{i=1}^{n} \rho(d_i; c) - b_0 \right],
\]

3
Let \( P_0 = [\psi_1^2, ..., \psi_p^2]^T \) and \( G_0 = [\gamma_1, ..., \gamma_q]^T \), an iterative system for the \( \gamma_j \) and \( \psi_j \) is given by

\[
P_0^{(s+1)} = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c)^{(t)}d_i^{(t)} \right]^{-1} C^{-1(t)}D^{(t)},
\]

\[
G_0^{(s+1)} = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c)^{(t)}d_i^{(t)} \right]^{-1} M^{-1(t)}N^{(t)},
\]

with the matrices \( C, D, M \) and \( N \) given in respectively (26), (27), (29) and (30). Note that \( \mu \) is calculated at each step with equation (6) and \( \Sigma \) from \( G_0 \) and \( P_0 \).

For the starting point of this iterative procedure, one can perform a confirmatory FA based on a robust “plug-in” covariance matrix that is fast to compute. Our choice goes to the orthogonocused Gnanadesikan-Kettering (OGK) proposed by Maronna and Zamar (2002). Alternatively, for an even faster solution, one can perform an exploratory FA (on e.g. the OGK) and set the resulting loadings (\( \gamma_{ij} \)) that are constrained (e.g. to zero) to the constrained value (e.g. zero). We successfully implemented this last alternative and used it for the simulations. The resulting algorithm is as follows:

1. Compute \( \mu_{start} = \mu_{OGK} \) and \( \Sigma_{OGK} \).
2. Compute \( \Psi_{start} \) and \( \Gamma_{start} \) by means of an exploratory factor analysis based on \( \Sigma_{OGK} \). Replace the fixed \( \gamma_{ij} \) in \( \Gamma_{start} \) by their value (e.g. zero).
3. Compute \( \Sigma_{start} = \Gamma_{start} \Sigma_{start} + \Psi_{start} \).
4. Compute the Mahalanobis distance \( d_i^{(1)} = \sqrt{(x_i - \mu_{start})^T \Sigma_{start}^{-1}(x_i - \mu_{start})} \).
5. Compute the different weights \( u(d_i^{(1)}; c) \).
6. Compute the mean vector

\[
\mu^{(1)} = \frac{\sum_{i=1}^{n} u(d_i^{(1)}; c)x_i}{\sum_{i=1}^{n} u(d_i^{(1)}; c)}.
\]

7. Compute the vector of loadings \( G_0 \) and uniqueness \( P_0 \)

\[
G_0^{(1)} = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c)^{(t)}d_i^{(t)} \right]^{-1} M^{-1(t)}N^{(t)},
\]

\[
P_0^{(1)} = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c)^{(t)}d_i^{(t)} \right]^{-1} C^{-1(t)}D^{(t)}.
\]

8. Using \( P_0^{(1)} \) and \( G_0^{(1)} \) compute \( \Sigma^{(1)} \).
9. Use a convergence criterion; if the conditions of convergence are met stop, otherwise start again from (4) and put \( \mu_{start} = \mu^{(1)} \) and \( \Sigma_{start} = \Sigma^{(1)} \). Repeat the procedure until convergence is reached.

3. Inference for the direct and two-stage estimators

### 3.1. Direct estimator

To compute the asymptotic covariance \( V_{G_0} \) of the loadings estimator \( G_0 \) and \( V_{P_0} \) of the variances estimator \( P_0 \), we use the results given by Lopuhaä (1989) for general \( S \)-estimators.
together with the delta-method. The asymptotic covariance of \(\sqrt{n}\text{vec}(\Sigma)\) is

\[
V_{\Sigma} = e_3(I_p' + K_{p,p})\Sigma \otimes \Sigma + e_4\text{vec}(\Sigma)\text{vec}(\Sigma)^T,
\]

where \(\otimes\) denotes the Kronecker product and \(K_{p,p}\) is a \((p^2 \times p^2)\)-block matrix with \((i,j)\)-block being a \(p \times p\) matrix with 1 at entry \((j,i)\) and 0 elsewhere. The constants \(e_3\) and \(e_4\) are given by

\[
e_3 = \frac{p(p + 2)E_\phi[\psi^2(d; c)d^2]}{E_\phi[\psi'(d; c)d^2 + (p + 1)\psi(d; c)d]},
\]

\[
e_4 = \frac{2}{p^3} + \frac{4E_\phi[\rho(d; c) - b_0]^2}{E_\phi[\psi(d; c)d]}.
\]

with \(p\) the dimension of \(\Sigma\), \(\psi(d; c) = (\partial/\partial d)p(d; c)\). The expectations are taken at the standardized \(p\)-variate normal distribution \(\Phi\). The values of \(e_3\) and \(e_4\) can be obtained by Monte-Carlo simulation.

The minimization of (7) provides the solutions \(\hat{G}_0\) and \(\hat{P}_0\) and \(\hat{\Sigma}\) is actually a function of the latter. To find the (asymptotic) variance of \(\hat{\gamma}\), we have

\[
\Gamma = \partial^2\hat{\gamma}/\partial \theta^2 = \frac{\partial}{\partial \theta}G_0 - \partial^2\psi(d; c)/\partial \theta^2 = \frac{\partial}{\partial \theta}(G_0 + \Psi(\hat{P}_0)).
\]

We then have

\[
V_{G_0} = (D_T\Gamma D_T)^{-1}D_TV_2D_T(D_T\Gamma D_T)^{-1},
\]

and

\[
V_{P_0} = (D_T\Gamma D_T)^{-1}D_TV_2D_T(D_T\Gamma D_T)^{-1}.
\]

For \(D_T\), we have

\[
D_T = \left[\frac{\partial}{\partial \gamma_1} \text{vec}(\Sigma) \ldots \frac{\partial}{\partial \gamma_q} \text{vec}(\Sigma)\right] = \left[\frac{\partial}{\partial \gamma_1} \text{vec}(\Gamma T^T) \ldots \frac{\partial}{\partial \gamma_q} \text{vec}(\Gamma T^T)\right],
\]

\[
= \left[\text{vec}\left(\frac{\partial}{\partial \gamma_1} \Gamma T^T\right) \ldots \text{vec}\left(\frac{\partial}{\partial \gamma_q} \Gamma T^T\right)\right],
\]

\[
= \left[\text{vec}\left(\frac{\partial}{\partial \gamma_1} \Gamma \Gamma^T\right) + \text{vec}(\Gamma^T \frac{\partial}{\partial \gamma_q} \Gamma) \ldots \right. \left. \text{vec}\left(\frac{\partial}{\partial \gamma_q} \Gamma \Gamma^T\right) + \text{vec}(\Gamma^T \frac{\partial}{\partial \gamma_q} \Gamma)\right].
\]

Because \(\text{vec}(AB) = (I_p \otimes A)\text{vec}(B) = (B^T \otimes I_m)\text{vec}(A)\) where \(A\) is an \(m \times n\) matrix and \(B\) is an \(n \times q\) matrix (see e.g. Magnus and Neudecker, 1988), we obtain that

\[
D_T = \left[\left(I_p \otimes \frac{\partial}{\partial \gamma_1} \Gamma\right) \text{vec}(\Gamma T^T) + \left(\frac{\partial}{\partial \gamma_1} \Gamma T^T \otimes I_p\right) \text{vec}(\Gamma) \ldots \right.
\]

\[
\left. \left(I_p \otimes \frac{\partial}{\partial \gamma_q} \Gamma\right) \text{vec}(\Gamma T^T) + \left(\frac{\partial}{\partial \gamma_q} \Gamma^T \otimes I_p\right) \text{vec}(\Gamma)\right].
\]

3.2. Two-stage estimator

The two-stage estimator \(\hat{\theta}\) is the result of the minimization of (3), i.e. the solution in \(\theta\) of

\[
\frac{\partial}{\partial \theta}W(\hat{\sigma} - \sigma(\theta)) = 0.
\]
The asymptotic variance of $\hat{\theta}$ is equal to (see e.g. Yuan and Bentler, 1998)

$$\Omega = A^{-1}\Pi A^{-1}, \quad (15)$$

with $A = \left(\frac{\partial \sigma(\theta)}{\partial \theta}\right)^T W \left(\frac{\partial \sigma(\theta)}{\partial \theta}\right)$, $\Pi = \left(\frac{\partial \sigma(\theta)}{\partial \theta}\right)^T W V_\Sigma W \left(\frac{\partial \sigma(\theta)}{\partial \theta}\right)^T$ and $V_\Sigma$ being the asymptotic variance of $\sqrt{n} \sqrt{n} \hat{\sigma} = \sqrt{n} \text{vec}(\hat{\Sigma})$. (14) actually defines a class of GLS estimators within which the most efficient (and equivalent to the MLE) is obtained when $W = V_\Sigma^{-1}$, which is replaced in practice by a sample estimate of $V_\Sigma$. Hence for the MLE, (15) simplifies to

$$\Omega = \left(\frac{\partial \sigma(\theta)}{\partial \theta}\right)^T V_\Sigma^{-1} \left(\frac{\partial \sigma(\theta)}{\partial \theta}\right)^{-1}. \quad (16)$$

When $\theta$ contains only the loadings, (16) can be written as

$$V_{G_0} = (D_T^T V_\Sigma^{-1} D_T)^{-1}, \quad (17)$$

with $V_\Sigma$ given in (8). One can see that (11) and (17) are equal if $D_T$ is a square matrix, but not in the other cases. It is however not clear which of the asymptotic covariance matrices is smaller, and hence which of the two-stage or direct estimators is the most efficient.

4. Simulation study

The aim of the simulation study is to compare the (direct) CBS estimator with respectively the classical two-stage estimator (i.e based on the Pearson correlations) and the two-stage estimator using the robust BS-estimator as covariance matrix estimator, under different kinds of models and data contamination. For both robust estimators, the breakdown point $\epsilon^*$ is set to 50%. For the classical and robust two-stage estimators we use the software Mplus, whereas we use the statistical package R (R Development Core Team, 2008) for the (direct) CBS, with the OGK estimator as a starting point as well as for computing the BS-estimator. We try both the MLE and the GLS objective function in Mplus for the classical and robust two-stage estimators, but the latter doesn’t seem to be reliable in relatively small samples ($n = 40, p = 20$) as will be shown below with Figure 1. Therefore, we will finally compare three estimators, the classical and robust two-stage estimators (with, respectively, the Pearson or the BS covariance estimates) with the ML objective function ($ML(P)$ and $ML(BS)$) and the (direct) CBS.

The data are generated from two different models, a model with a relatively small number of manifest variables ($p = 6$) for which we simulate relatively large samples ($n = 100$) and a model with a relatively large number of manifest variables ($p = 20$) for which we simulate relatively small samples ($n = 40$). In both models, we set the number of latent variables to 2, and for each of them, two sets of values for $\Gamma$ were chosen in order to create a strong and weak signal to noise ratio. The values of the parameters for all models can be found in Appendix B. For the model with $p = 6$, the number of loadings different from 0 ($q$) is equal to $p$ and for the other model we choose 2 different values for $q$, namely $q = p = 20$ and $q = 26$.

The contamination consists in generating $(1 - \epsilon\%)$ of the data from a multivariate normal distribution with parameters given by the model and $\epsilon\%$ of the data from a multivariate normal distribution with one third of the loadings different from 0 that are multiplied by 10. Different amounts of contamination are chosen, but we present here only the $\epsilon = 5\%$ case.

For each model, we compute the asymptotic efficiency measured by the ratio of the traces of the covariance matrices of the robust two-stage relative to the direct CBS estimators using (11)
and (17) with the true values of the parameters. With our chosen set of parameters, when $q = p$ the efficiency is approximately one but when $q > p$ the two-stage estimator is asymptotically more efficient (see values in Appendix B).

In Figure 1 are presented the bias distributions of the loadings’ and uniqueness’ estimators for the model with $n = 40$ ($q = p = 20$) and a high signal to noise ratio, without data contamination. The estimators are the two-stage estimators using respectively the $ML$ or the $GLS$ objective functions, with the Pearson correlations or the $BS$ estimator, as well as the direct $CBS$ estimation. As mentioned previously, it seems that using the $GLS$ objective function produces unreliable results, while the other estimators ($ML(P), ML(BS)$ and $CBS$) behave well (unbiased) and in the same manner (same variances).

When there is data contamination, not surprisingly the simulations show that the two-stage estimator based on the Pearson’s covariance matrix is biased, whereas this is not the case with the $CBS$ and the two-stage estimator based on the $BS$ (see Figure 2).

Overall, the two robust estimators ($ML(BS)$ and $CBS$) seem to be pretty similar in terms of bias and variance. However, in one case ($n = 100, p = 6$ and low signal to noise ratio) the two-stage estimator produces extreme estimates for the loadings and negative ones for the uniqueness, as can be seen in Figures 3 and 4. The Mean Square Error of the two-stage estimator is really larger than that of the $CBS$ in that case (see Table 1). This could be due to the fact that the two-stage estimator needs the estimation of a full covariance matrix.

In the case where $q > p$, although the asymptotic variance of the $CBS$ is larger than that
of the two-stage estimator, with finite samples, the CBS doesn’t seem to be more variable than the two-stage estimator with or without data contamination, as illustrated in Figure 5. This might be explained by the fact that since $p$ is large relative to $n$, the BS estimator can be very (computationally) unstable so that its finite sample variance is (artificially) increased. With the CBS the number of parameters to estimate is smaller relative to the sample size, hence providing more (computationally) stable estimators.

To conclude, the simulation study shows that, in finite samples, the CBS is not less efficient than the indirect estimator and it is more stable in some cases where the latter produces extreme estimates.
Figure 3: Bias distribution of the loadings’ and uniqueness’ estimators for the model with $p = 6$ without data contamination and with a low signal to noise ratio.

Table 1: MSE $\times 100$ for the model with $p = 6$, low signal to noise ratio

<table>
<thead>
<tr>
<th>$\epsilon^a$ = 0%</th>
<th>$\epsilon^a$ = 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$</td>
<td>ML(P)</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>5.45</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>9.35</td>
</tr>
<tr>
<td>$\gamma_4$</td>
<td>10.63</td>
</tr>
<tr>
<td>$\gamma_5$</td>
<td>1.24</td>
</tr>
<tr>
<td>$\gamma_6$</td>
<td>6.86</td>
</tr>
<tr>
<td>$\psi_1$</td>
<td>13.45</td>
</tr>
<tr>
<td>$\psi_2$</td>
<td>174.50</td>
</tr>
<tr>
<td>$\psi_3$</td>
<td>56.35</td>
</tr>
<tr>
<td>$\psi_4$</td>
<td>62.63</td>
</tr>
<tr>
<td>$\psi_5$</td>
<td>1.51</td>
</tr>
<tr>
<td>$\psi_6$</td>
<td>30.47</td>
</tr>
</tbody>
</table>

$^a$Amount of contamination
Figure 4: Bias distribution of the loadings' and uniqueness' estimators for the model with $p = 6$ with data contamination and with a low signal to noise ratio.
Figure 5: Bias distribution of the loadings’ and uniqueness’ estimators for the model with $n = 40, q > p$, data contamination and with a low signal to noise ratio.
5. Conclusion

To estimate in a robust fashion the parameters of a FA model, one can choose between two strategies. One can compute a two-stage estimator by which a robust covariance matrix of the manifest variables is first computed and then used in the MLE or GLS estimating equations, possibly using a software, such as Mplus, available in the market. Alternatively, one can directly estimate the FA parameters (loadings and uniqueness) from a robust objective function such as the one of an $S$-estimator, in the same spirit as is done in Copt and Victoria-Feser (2006) for the mixed linear model. In this paper we derive the estimating equations (and an iterative procedure) for the $S$-estimator of the FA model, i.e. the CBS-estimator. The two estimators have different asymptotic properties, in that their asymptotic covariance matrix is not the same. With the models used in the simulation study, we found that the (trace of the) asymptotic covariance matrix of the two-stage estimator is even smaller. However, in finite samples, the CBS-estimator seems to be more stable than the two-stage estimator (smaller MSE).

Finally, the aim of this paper was to compare different approaches to robust estimation of FA models. A step further would be to evaluate the models using a goodness-of-fit type criterion or test statistic. A natural choice is a likelihood ratio type test on the covariance matrix (see e.g. Bartholomew and Knott, 1999) which compares the log-likelihood evaluated with $\Sigma = \Gamma(G_0)\Gamma^T(G_0) + \Psi(P_0)$ the constrained covariance matrix and that evaluated with the unconstrained covariance matrix. The results on robust classes of test of Heritier and Ronchetti (1994) could be used with some probable adaptations. We however leave this problem for future research.
A. Derivation of the estimating equations for the CBS-estimator

Using \( \frac{\partial}{\partial \theta} \log |\Sigma| = \text{tr} \left[ \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta} \right] \) and \( \frac{\partial}{\partial \theta} \Sigma^{-1} = -\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta} \Sigma^{-1} \) with \( \theta_j \) any element of \( \theta \), the vector of the FA parameters, as well as \( \frac{\partial}{\partial \gamma} \Sigma = \left( \frac{\partial \Gamma}{\partial \gamma} \right) \Gamma^T + \Gamma \left( \frac{\partial \Gamma^T}{\partial \gamma} \right) \) the partial derivatives of the Lagrangian (7) are then

\[
\frac{\partial}{\partial \mu} L = -\lambda_1 \frac{1}{n} \sum_{i=1}^{n} u(d_i; c) \Sigma^{-1} (x_i - \mu) = 0, \tag{18}
\]

\[
\frac{\partial}{\partial \psi_j} L = \text{tr} \left[ \Sigma^{-1} Z_j \right] - \lambda_2 \frac{1}{n} \sum_{i=1}^{n} u(d_i; c)(x_i - \mu)^T \Sigma^{-1} Z_j \Sigma^{-1} (x_i - \mu) = 0, \tag{19}
\]

with \( Z_j = \frac{\partial}{\partial \gamma_j} \Sigma = \frac{\partial}{\partial \gamma_j} (\Gamma \Gamma^T + \Psi) \), and

\[
\frac{\partial}{\partial \gamma_j} L = \text{tr} \left[ \Sigma^{-1} \left( \frac{\partial}{\partial \gamma_j} \Gamma \right) \Gamma^T + \Sigma^{-1} \left( \frac{\partial}{\partial \gamma_j} \Gamma^T \right) \right] - \lambda_2 \frac{1}{n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} \left( \frac{\partial}{\partial \gamma_j} \Gamma \right) \Gamma^T \Sigma^{-1} (x_i - \mu) = 0. \tag{20}
\]

From (18) we get the estimating equation for \( \mu \) given in (6). Then, multiplying (19) by \( \psi_j^2 \) and using the properties of the trace we get

\[
\text{tr} \left[ \Sigma^{-1} Z_j \psi_j^2 \right] - \lambda_2 \frac{1}{2n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} Z_j \psi_j^2 \Sigma^{-1} (x_i - \mu) = 0,
\]

\[
\sum_{j=1}^{r} \text{tr} \left[ \Sigma^{-1} Z_j \psi_j^2 \right] - \lambda_2 \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{r} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} Z_j \psi_j^2 \Sigma^{-1} (x_i - \mu) = 0,
\]

\[
\text{tr} \left[ \Sigma^{-1} \sum_{j=1}^{r} Z_j \psi_j^2 \right] - \lambda_2 \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{r} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} \sum_{j=1}^{r} Z_j \psi_j^2 \Sigma^{-1} (x_i - \mu) = 0,
\]

\[
\text{tr} \left[ \Sigma^{-1} \Psi \right] - \lambda_2 \frac{1}{2n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} \Psi \Sigma^{-1} (x_i - \mu) = 0. \tag{21}
\]
Multiplying (21) by 2 and adding the result to (22) we obtain

\[
\text{so that we finally get}
\]

\[
\lambda \sum_{j=1}^{n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T (A_{ij}) (A_{ij}) \gamma_j - 
\]

\[
\sum_{j=1}^{q} \text{tr} \left[ \left( \sum_{i=1}^{n} (A_{ij}) \gamma_j \right) (A_{ij}) \right] = 0,
\]

\[
\lambda \sum_{j=1}^{n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T (A_{ij}) (A_{ij}) \gamma_j - 
\]

\[
\sum_{j=1}^{q} \text{tr} \left[ \left( \sum_{i=1}^{n} (A_{ij}) \gamma_j \right) (A_{ij}) \right] = 0,
\]

\[
2 \text{tr} \left[ \Sigma^{-1} \Sigma \right] \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} (A_{ij}) (A_{ij}) \gamma_j = 0. \quad (22)
\]

Multiplying (21) by 2 and adding the result to (22) we obtain

\[
2 \text{tr} \left[ \Sigma^{-1} \gamma_j \right] + 2 \text{tr} \left[ \Sigma^{-1} \gamma_j \right] = \frac{\lambda}{n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} \gamma_j \gamma_j = 0,
\]

\[
\lambda \sum_{j=1}^{n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T (A_{ij}) (A_{ij}) \gamma_j = 0,
\]

\[
2 \text{tr} \left[ \Sigma^{-1} \gamma_j \right] + 2 \text{tr} \left[ \Sigma^{-1} \gamma_j \right] = \frac{\lambda}{n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T (\gamma_j \gamma_j) \gamma_j = 0,
\]

\[
2 \text{tr} \left[ \gamma_j \gamma_j \right] - \frac{\lambda}{n} \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} (A_{ij}) (A_{ij}) \gamma_j = 0.
\]

so that we finally get \( \lambda = 2p \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c)(d_i^2) \right] \). Using this \( \lambda \) in (19) and (20) respectively yields

\[
\text{tr} \left[ \Sigma^{-1} \gamma_j \right] = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c)(d_i^2) \right]^{-1} \left[ \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} (A_{ij}) (A_{ij}) \gamma_j \right]. \quad (23)
\]
and

\[
\operatorname{tr} \left[ \Sigma^{-1} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right)^T + \Sigma^{-1} \frac{\partial \Gamma}{\partial \gamma_j} \right] = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c) d_i^T \right],
\]

(24)

\[
\sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right)^T (\Gamma^T + \frac{\partial \Gamma}{\partial \gamma_j}) \Sigma^{-1} (x_i - \mu).
\]

(23) and (24) are the estimating equations for the parameters \( \gamma_j \) and \( \psi_j \).

To define an iterative system, we need now to extract \( \psi_j \) and \( \gamma_j \) from (23) and (24). First, we use \( \Psi = \sum_{j=1}^{r} Z_j \psi_j \) to reformulate the left hand side of (23) as

\[
\operatorname{tr} \left[ \Sigma^{-1} Z_j \right] = \operatorname{tr} \left[ \Sigma^{-1} Z_j \Psi \right] = \operatorname{tr} \left[ \Sigma^{-1} Z_j \Psi \right] \sum_{k=1}^{r} Z_k \psi_k^2,
\]

(25)

which defines an iterative system for each \( \psi_k^2, k = 1, \ldots, r \). Let

\[
C = \left[ \operatorname{tr} \left[ \Sigma^{-1} Z_j \Psi \right] \right]_{j=k=1, \ldots, r},
\]

and

\[
D = \left[ \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} Z_j \Sigma^{-1} (x_i - \mu) \right]_{j=1, \ldots, r},
\]

(27)

an expression for the \( \psi_j^2 \) is given by

\[
P_0 = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c) d_i^T \right]^{-1} C^{-1} D.
\]

(28)

In the same way, we use \( \Gamma = \sum_{j=1}^{q} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right) \gamma_j \) to reformulate the left hand side of (24) as

\[
\operatorname{tr} \left[ \Sigma^{-1} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right)^T \right] + \Sigma^{-1} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right) = \operatorname{tr} \left[ \Sigma^{-1} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right) + \Sigma^{-1} \frac{\partial \Gamma}{\partial \gamma_j} \right] \left( \frac{\partial \Gamma}{\partial \gamma_j} \right) \gamma_k.
\]

(30)

Letting

\[
M = \left[ \operatorname{tr} \left[ \Sigma^{-1} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right) \right] \right]_{j,k=1, \ldots, q},
\]

(29)

and

\[
N = \left[ \sum_{i=1}^{n} u(d_i; c) (x_i - \mu)^T \Sigma^{-1} \left( \frac{\partial \Gamma}{\partial \gamma_j} \right)^T \Sigma^{-1} (x_i - \mu) \right]_{j=1, \ldots, q},
\]

so that

\[
G_0 = \frac{p}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} u(d_i; c) d_i^T \right]^{-1} M^{-1} N.
\]

(31)
B. Parameter’s values for the simulations

B.1. First Model : \( n = 100, \ p = 6 \)

We first choose the parameter’s values \( \mu, \ \Gamma \) and \( \Psi \) so that the signal to noise ratio is high. They are

\[
\mu = [0, 0, 0, 0, 0, 0]^T, \quad \Gamma = \begin{bmatrix}
-1.5 & 0 \\
1.8 & 0 \\
2 & 0 \\
0 & 1.93 \\
0 & -1.6 \\
0 & 1.4 \\
\end{bmatrix}, \quad \Psi = I_p.
\]

The efficiency of the ML(\( BS \)) relative to the CBS is 0.996.

For a low signal to noise ratio case, we choose for \( \mu, \ \Gamma \) and \( \Psi \) the following values

\[
\mu = [0, 0, 0, 0, 0, 0]^T, \quad \Gamma = \begin{bmatrix}
0.1 & 0 \\
0.54 & 0 \\
-0.45 & 0 \\
0 & 0.83 \\
0 & 0.26 \\
0 & -0.6 \\
\end{bmatrix}, \quad \Psi = I_p.
\]

The efficiency of the ML(\( BS \)) relative to the CBS is 0.978.

B.2. Second Model : \( n = 40, \ p = 20 \)

For a high signal to noise ratio and \( p = q \), the values for \( \mu, \ \Gamma \) and \( \Psi \) are

\[
\mu = [0, \ldots, 0]^T, \quad \Gamma = \begin{bmatrix}
-1.2 & 0 \\
-1.3 & 0 \\
-1.6 & 0 \\
1.3 & 0 \\
0.9 & 0 \\
2.2 & 0 \\
1.2 & 0 \\
2.1 & 0 \\
1.5 & 0 \\
0.3 & 0 \\
0 & -1.1 \\
0 & 1.5 \\
0 & -2.1 \\
0 & -0.7 \\
0 & 2.2 \\
0 & 1.5 \\
0 & -2 \\
0 & 0.9 \\
0 & 0.7 \\
0 & -0.8 \\
\end{bmatrix}, \quad \Psi = I_p.
\]
The efficiency of the $ML(BS)$ relative to the $CBS$ is 0.999.

For the low signal to noise ratio, the values for $\mu$, $\Gamma$ and $\Psi$ are

$$
\mu = [0, \ldots, 0]^T, \quad \Gamma = \begin{bmatrix}
-0.7 & 0 \\
-0.4 & 0 \\
0.1 & 0 \\
-0.6 & 0 \\
-0.62 & 0 \\
-0.43 & 0 \\
0.4 & 0 \\
1 & 0 \\
0.47 & 0 \\
-0.3 & 0 \\
0 & 0.1 \\
0 & 0.8 \\
0 & -0.6 \\
0 & 0.5 \\
0 & -0.9 \\
0 & -0.95 \\
0 & -0.8 \\
0 & 0.3 \\
0 & 0.03 \\
0 & 0.09 
\end{bmatrix}, \quad \Psi = I_p.
$$

The efficiency of the $ML(BS)$ relative to the $CBS$ is 0.999.

For a high signal to noise ratio and 26 loadings different from zero, the values for $\mu$, $\Gamma$ and $\Psi$
are

\[
\begin{bmatrix}
-1.2 & 0 \\
-1.3 & 0 \\
-1.6 & -1.2 \\
1.3 & 0 \\
0.9 & -0.9 \\
2.2 & 0 \\
1.2 & 0 \\
2.1 & 0 \\
1.5 & 1.4 \\
0.3 & 0 \\
0 & -1.1 \\
0 & 1.5 \\
1.3 & -2.1 \\
0 & -0.7 \\
0 & 2.2 \\
0 & 1.5 \\
-1 & -2 \\
1.6 & 0.9 \\
0 & 0.7 \\
0 & -0.8
\end{bmatrix}, \quad \Psi = I_p.
\]

The efficiency of the \(ML(\hat{S})\) relative to the \(CBS\) is 0.832.

For a low signal to noise ratio and 26 loadings different from zero, the values for \(\mu\), \(\Gamma\) and \(\Psi\) are

\[
\mu = [0, \ldots, 0]^T, \quad \Gamma = \begin{bmatrix}
-0.7 & 0 \\
-0.4 & 0 \\
0.1 & -0.2 \\
-0.6 & 0 \\
-0.62 & -0.45 \\
-0.43 & 0 \\
0.4 & 0 \\
1 & 0 \\
0.47 & 0.31 \\
-0.3 & 0 \\
0 & 0.1 \\
0 & 0.8 \\
0.3 & -0.6 \\
0 & 0.5 \\
0 & -0.9 \\
0 & -0.95 \\
-0.8 & -0.8 \\
0.64 & 0.3 \\
0 & 0.03 \\
0 & 0.09
\end{bmatrix}, \quad \Psi = I_p.
\]

The efficiency of the \(ML(\hat{S})\) relative to the \(CBS\) is 0.952.
References


