

# Factor analysis and outliers: A Bayesian approach

Wolfgang Polasek

Institute of Statistics and Econometrics  
University of Basel  
Holbeinstrasse 12, 4051 Basel, Switzerland  
Email: Wolfgang@iso.iso.unibas.ch  
February 24, 2000

## Abstract

Classical factor analysis decomposes  $n$  observations of dimension  $p$  into  $K$  ( $<p$ ) orthogonal factors. In a Bayesian approach we decompose the observation matrix into a product of a factor score and a factor loading matrix of unknown rank by using a normal-Wishart conjugate density family. We assume an informative prior and show how the posterior distribution can be simulated in multivariate blocks by a Gibbs sampling algorithm. The number of factors is determined using the ordinary marginal likelihood and the posterior marginal likelihood criteria.

Furthermore, the sensitivity of the factor analysis with respect to outliers in the data set is explored. Assuming additive outliers, a Gibbs sampling approach is suggested for a multivariate outlier model in extension of the approach of Verdinelli and Wasserman (1991). The approach is demonstrated for the language data set of Fuller (1987).

**Keywords:** Factor analysis, Gibbs sampling, marginal likelihood, multivariate outliers.

# 1 Introduction

Factor analysis is a popular tool in social sciences (e.g., economics, psychology and sociology) for reducing a multivariate data matrix to a matrix product of rank lower than the number of variables. This is particularly useful if one assumes that many variables are measuring similar things which can be summarized by some few ‘latent’ or factor variables. There are only a few Bayesian approaches to factor analysis, like Press and Shigemasu (1989), and Martin and McDonald (1975), who had to use rather restrictive model assumptions since at that time the numerical techniques for multivariate Bayesian data analysis were rather limited. The question arises as to whether or not the new technique of Monte Carlo Markov Chain (MCMC) methods can be used to advance Bayesian inference in factor analysis. In this paper we propose a coherent approach to fit a Bayesian factor analysis which can take into account possible multivariate outliers. Extending the Gibbs sampling approach of Verdinelli and Wasserman (1991) for multivariate location shift outlier models, we show how to derive the conditional distributions for the Gibbs sampling algorithm.

Thus, the marginal likelihood can be used to determine the number of factors in a factor analysis model and also to test for outliers. The ratio of marginal likelihoods defines the Bayes factors and we show how the approach of Chib (1995) can be used to compute the marginal likelihoods from Gibbs sampling output. In the appendix we show how the full conditional distributions of the factor analysis model can be derived in closed form. Previously, a parsimonious approach to factor analysis using the information criterion AIC was suggested by Akaike (1987).

Though various approaches exist to create outliers in univariate models (as in Kitagawa and Akaike (1982), Barnett and Lewis (1984), Pettit and Smith (1985)), little work has been done for outlier models in multivariate Bayesian analysis. Therefore we will suggest the simple location shift outlier model as a basic model for detecting outliers in factor analysis. In univariate comparisons, location shift outlier models have been found superior to, e.g., multiplicative or variance inflation outliers because they produce more likely aberrant observations than “inliers”. Also Rocke and Woodruff (1996) show that multivariate outliers with local shift and the same covariance matrix as the ‘good’ observations are the most difficult ones to detect. Fortunately from Bayesian point of view location shift outliers can be easily handled by

the Gibbs sampler (see Verdinelli and Wassermann (1991)). Any classical or Bayesian method to determine outliers in a multivariate model is computational burdensome. But the proposed Gibbs sampling approach has the advantage that it covers the largest set of models by a rather simple estimation strategy. The number of factors and the presence of outliers can be detected from computing the marginal likelihoods at the same time.

The plan of the paper is as follows: In section two we introduce the Bayesian model of factor analysis. The marginal likelihood for model choice and the number of factors is calculated in order to determine the number of factors. In section three, we discuss the factor analysis model with outliers. This is a multivariate extension of the outlier approach of Verdinelli and Wasserman (1991). In section four we analyse the language data set of Fuller (1987) and describe inferences for possible (location shift) outliers. Section five concludes and in the appendix we have listed computational details for the full conditional distribution of the Gibbs sampler and the components of the marginal likelihood.

## 2 Bayesian Factor Analysis

We start with a factor analysis model where the number of factors  $K$  is known. Let  $\mathbf{y}_1, \dots, \mathbf{y}_n$  be  $n$  observations from a  $p$ -dimensional normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\mathbf{\Psi}$ . The factor analysis model addresses the question if the  $p$ -dimensional observations of factors  $\mathbf{y}_1, \dots, \mathbf{y}_n$  can be reduced to  $n$   $K$ -dimensional latent observations  $\mathbf{z}_1, \dots, \mathbf{z}_n$ , i.e.

$$\mathbf{y}_i = \mathbf{\Lambda}\mathbf{z}_i + \mathbf{e}_i, \quad i = 1, \dots, n, \quad (1)$$

and a  $p \times K$  dimensional matrix  $\mathbf{\Lambda}$  of factor loadings of full rank. The latent factors  $\mathbf{z}_i$  are assumed to be normally independently distributed  $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Phi})$  and are also independently distributed from the error term  $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$ . For the Bayesian factor analysis we specify a normal prior distribution for  $\mathbf{\Lambda}$  and a Wishart prior distribution for covariance matrices  $\mathbf{\Psi}$  and  $\mathbf{\Phi}$  (see below). While classical factor analysis sets  $\mathbf{\Phi} = \mathbf{I}$  and uses a diagonal  $\mathbf{\Psi}$  matrix, we impose these restrictions via the prior information matrices  $\mathbf{\Psi}_*$  and  $\mathbf{\Phi}_*$ . Since we assume informative prior distributions throughout the paper, all hyperparameters of the prior distributions will be denoted by symbols



The parameters of the factor analysis are given in four blocks, i.e. by  $\theta = (\mathbf{\Lambda}, \mathbf{Z}, \mathbf{\Psi}, \mathbf{\Phi})$  and  $\mathbf{\Lambda}_*, \mathbf{H}_*, \mathbf{G}_*, \mathbf{\Psi}_*, n_*, \mathbf{\Phi}_*$  and  $\nu_*$ , are known hyperparameters of the prior distributions and the joint distribution of the data  $\mathbf{Y}$  and the parameter  $\theta$  is

$$\begin{aligned}
p(\mathbf{Y}, \theta) &= \mathcal{N}[\text{vec } \mathbf{Y} \mid \text{vec } \mathbf{\Lambda}\mathbf{Z}, \mathbf{I}_n \otimes \mathbf{\Psi}] \cdot \mathcal{N}[\text{vec } \mathbf{Z} \mid \mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Phi}] \cdot \\
&\quad \cdot \mathcal{N}[\text{vec } \mathbf{\Lambda} \mid \mathbf{\Lambda}_*, \mathbf{H}_* \otimes \mathbf{G}_*] \cdot \mathcal{W}_p[\mathbf{\Psi}^{-1} \mid \mathbf{\Psi}_*, n_*] \cdot \mathcal{W}_K[\mathbf{\Phi}^{-1} \mid \mathbf{\Phi}_*, \nu_*] \\
&\propto |\mathbf{I}_n \otimes \mathbf{\Psi}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \text{vec}'(\mathbf{Y} - \mathbf{\Lambda}\mathbf{Z})'(\mathbf{I}_n \otimes \mathbf{\Psi}^{-1}) \text{vec}(\mathbf{Y} - \mathbf{\Lambda}\mathbf{Z})\right\} \cdot \\
&\quad |\mathbf{I}_n \otimes \mathbf{\Phi}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \text{vec}'\mathbf{Z}'(\mathbf{I}_n \otimes \mathbf{\Phi}^{-1}) \text{vec}\mathbf{Z}\right\} \cdot |\mathbf{H}_* \otimes \mathbf{G}_*|^{-\frac{1}{2}} \cdot \\
&\quad \exp\left\{-\frac{1}{2} \text{vec}'(\mathbf{\Lambda} - \mathbf{\Lambda}_*)(\mathbf{H}_* \otimes \mathbf{G}_*)^{-1} \text{vec}(\mathbf{\Lambda} - \mathbf{\Lambda}_*)\right\} \cdot \\
&\quad |\mathbf{\Psi}|^{-\frac{n_*-p-1}{2}} \exp\left\{-\frac{1}{2} \text{tr}\mathbf{\Psi}^{-1}\mathbf{\Psi}_*\right\} \cdot |\mathbf{\Phi}|^{-\frac{\nu_*-K-1}{2}} \exp\left\{-\frac{1}{2} \text{tr}\mathbf{\Phi}^{-1}\mathbf{\Phi}_*\right\}. \quad (4)
\end{aligned}$$

From the joint distribution, the full conditional distributions (f.c.d.) can be derived in the following way:

1. The f.c.d. for the latent variables  $\mathbf{Z}$  is a multivariate normal distribution

$$p(\mathbf{Z} \mid \theta^c, \mathbf{Y}) = \mathcal{N}[\mathbf{Z}_{**}, \mathbf{D}_Z] \quad (5)$$

with the parameters

$$\begin{aligned}
\mathbf{D}_Z^{-1} &= \mathbf{I}_n \otimes \mathbf{\Phi}^{-1} + \mathbf{I}_n \otimes \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}, \\
\text{vec } \mathbf{Z}_{**} &= \mathbf{D}_Z \text{vec}(\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{Y})
\end{aligned} \quad (6)$$

or, in matrix notation,

$$\mathbf{Z}_{**} = (\mathbf{\Phi}^{-1} + \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{Y}. \quad (7)$$

We have used the following relation to vectorize a product of matrices (for the vec-operator see e.g., Magnus and Neudecker (1988)):

$$\text{vec } \mathbf{\Lambda}\mathbf{Z} = (\mathbf{I}_n \otimes \mathbf{\Lambda}) \text{vec } \mathbf{Z}.$$

2. The f.c.d. for the factor loading matrix

$$p(\mathbf{\Lambda} \mid \theta^c, \mathbf{Y}) = \mathcal{N}[\mathbf{\Lambda}_{**}, \mathbf{C}_{**}] \quad (8)$$

is a multivariate normal distribution with

$$\begin{aligned} \mathbf{C}_{**}^{-1} &= \mathbf{H}_*^{-1} \otimes \mathbf{G}_*^{-1} + \mathbf{Z}\mathbf{Z}' \otimes \mathbf{\Psi}^{-1}, \\ \text{vec } \mathbf{\Lambda}_{**} &= \mathbf{C}_{**}[(\mathbf{H}_*^{-1} \otimes \mathbf{G}_*^{-1})\text{vec } \mathbf{\Lambda}_* + (\mathbf{Z} \otimes \mathbf{\Psi}^{-1})\text{vec } \mathbf{Y}] \\ &= \mathbf{C}_{**}[\text{vec } (\mathbf{G}_*^{-1} \mathbf{\Lambda}_* \mathbf{H}_*^{-1} + \mathbf{\Psi}^{-1} \mathbf{Y}\mathbf{Z})]. \end{aligned}$$

3. For the f.c.d. for the variance matrix  $\mathbf{\Psi}$ , we find a Wishart distribution

$$p(\mathbf{\Psi}^{-1} \mid \theta^c, \mathbf{Y}) = \mathcal{W}_p[\mathbf{\Psi}_{**}, n_{**} = n_* + n] \quad (9)$$

with the parameter

$$\mathbf{\Psi}_{**} = \mathbf{\Psi}_* + (\mathbf{Y} - \mathbf{\Lambda}\mathbf{Z})(\mathbf{Y} - \mathbf{\Lambda}\mathbf{Z})'.$$

4. The f.c.d. for the variance matrix  $\mathbf{\Phi}$  is also a Wishart distribution

$$p(\mathbf{\Phi}^{-1} \mid \theta^c, \mathbf{Y}) = \mathcal{W}_K[\mathbf{\Phi}_{**}, \nu_{**} = \nu_* + n] \quad (10)$$

with

$$\mathbf{\Phi}_{**} = \mathbf{\Phi}_* + \mathbf{Z}\mathbf{Z}'.$$

Note that last step can be skipped if we assume a spherical normal distribution for the latent variables  $\mathbf{Z}$  and set  $\mathbf{\Phi} = \mathbf{I}_k$  in (6) and (7).

## 2.1 Selecting the number of factors

In order to choose between factor analysis models with different numbers of factors (in short: factor analysis of rank  $r$ ), we use the Bayes factor and the marginal likelihood as model choice criteria (see also section 3.2). For equal prior probabilities the Bayes factor is given by the ratio of marginal likelihoods

$$B = \frac{p(\mathbf{Y} \mid \text{rank} = r)}{p(\mathbf{Y} \mid \text{rank} = s)} \quad (11)$$

and can be used to choose between factor analysis models with rank  $r$  or  $s$ . The marginal likelihood for a certain rank specification is calculated by

$$p(\mathbf{Y}) = \frac{p(\mathbf{Y} | \hat{\theta}) p(\hat{\theta})}{p(\hat{\theta} | \mathbf{Y})}. \quad (12)$$

For the denominator in (12) we need the following posterior ordinate decomposition for an appropriate chosen point  $\hat{\theta} = (\hat{\mathbf{Z}}, \hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}})$ :

$$p(\hat{\theta} | \mathbf{Y}) = p(\hat{\mathbf{Z}} | \mathbf{Y}) \times p(\hat{\mathbf{\Lambda}} | \hat{\mathbf{Z}}, \mathbf{Y}) \times p(\hat{\mathbf{\Psi}}^{-1} | \hat{\mathbf{\Lambda}}, \hat{\mathbf{Z}}, \mathbf{Y}) \times p(\hat{\mathbf{\Phi}}^{-1} | \hat{\mathbf{\Psi}}^{-1}, \hat{\mathbf{\Lambda}}, \hat{\mathbf{Z}}, \mathbf{Y}). \quad (13)$$

We show in Appendix B how the four components in (13) can be computed by additional steps in the Gibbs algorithm. Following Chib (1995), the three components on the right hand side of (12), which depend on  $\hat{\theta}$ , are computed in the following way:

The likelihood ordinate in (12) is calculated by

$$p(\mathbf{Y} | \hat{\theta}) = \mathcal{N}_{p \times n}[\mathbf{Y} | \hat{\mathbf{\Lambda}} \hat{\mathbf{Z}}, \mathbf{I}_n \otimes \hat{\mathbf{\Psi}}] \quad (14)$$

and the ordinate of the prior density is given by

$$p(\hat{\theta}) = \mathcal{W}[\hat{\mathbf{\Psi}}^{-1} | \mathbf{\Psi}_*, n_*] \cdot \mathcal{W}[\hat{\mathbf{\Phi}}^{-1} | \mathbf{\Phi}_*, \nu_*] \cdot \mathcal{N}[\hat{\mathbf{Z}} | \mathbf{0}, \mathbf{I}_n \otimes \hat{\mathbf{\Phi}}] \cdot \mathcal{N}[\hat{\mathbf{\Lambda}} | \mathbf{\Lambda}_*, \mathbf{H}_* \otimes \mathbf{G}_*]. \quad (15)$$

Alternatively we can compute the logarithm of the marginal likelihood (12) by

$$\log p(\mathbf{Y}) = \log p(\mathbf{Y} | \hat{\theta}) + \log p(\hat{\theta}) - \log p(\hat{\theta} | \mathbf{Y}). \quad (16)$$

This (log) marginal likelihood can be calculated for models with different number of factors  $k = 1, \dots, p$  and we choose the model with the highest marginal likelihood.

### 3 Factor analysis with outliers

Consider the factor analysis model in (1) as a sample from the normal distribution

$$\mathbf{y}_i \sim \mathcal{N}_p[\mathbf{\Lambda} \mathbf{z}_i, \mathbf{\Psi}], \quad i = 1, \dots, n,$$

which is expose to occasional additive (or location shift) outliers. The univariate location shift outlier model was first analyzed by the Gibbs sampler

in Verdinelli and Wasserman (1991). The multivariate location shift outlier model is formulated with the  $n$  indicator variables  $\vartheta_1, \dots, \vartheta_n$

$$\begin{aligned} f(\mathbf{y}_i | \vartheta_i) &= (1 - \vartheta_i) \mathcal{N}_p[\mathbf{y}_i | \mathbf{\Lambda} \mathbf{z}_i, \mathbf{\Psi}] \\ &+ \vartheta_i \mathcal{N}_p[\mathbf{y}_i | \mathbf{a}_i + \mathbf{\Lambda} \mathbf{z}_i, \mathbf{\Psi}] \end{aligned} \quad (17)$$

where  $\mathbf{y}_i$  is the  $i$ -th column of  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ , and  $\mathbf{a}_i$  is the  $i$ -th column of  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ . We assume that each indicator  $\vartheta_i$  is distributed as a Bernoulli random variable with parameter  $\varepsilon_*$ , the prior probability that the  $i$ -th observation is an outlier. Let  $\mathbf{D}_\vartheta = \text{diag}(\vartheta_1, \dots, \vartheta_n)$  be a  $n \times n$  indicator matrix for the (multivariate) outliers. Then the factor analysis model with outliers can be written as multivariate regression model

$$\mathbf{Y} \sim \mathcal{N}_{p \times n}[\mathbf{\Lambda} \mathbf{Z} + \mathbf{A} \mathbf{D}_\vartheta, \mathbf{I}_n \otimes \mathbf{\Psi}] \quad (18)$$

with the prior information given compactly as

$$\begin{aligned} \mathbf{Z} &\sim \mathcal{N}_{K \times n}[\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Phi}_*], \\ \mathbf{\Lambda} &\sim \mathcal{N}_{p \times K}[\mathbf{\Lambda}_*, \mathbf{H}_* \otimes \mathbf{G}_*], \\ \mathbf{\Psi} &\sim \mathcal{W}_p[\mathbf{\Psi}_*, n_*], \quad \mathbf{\Phi} \sim \mathcal{W}_k[\mathbf{\Phi}_*, \nu_*], \\ \mathbf{A} &\sim \mathcal{N}_{p \times n}[\mathbf{A}_*, \mathbf{I}_n \otimes \mathbf{P}_*], \\ \vartheta_i &\sim \text{Ber}[\varepsilon_*], \quad i = 1, \dots, n, \end{aligned} \quad (19)$$

where  $\mathbf{A}_* : p \times K$  and  $\mathbf{P}_* : p \times p$  are a-priori known parameter matrices for the location and variances of outliers and  $\text{Ber}[\varepsilon_*]$  denotes a Bernoulli distributed random variable with known success probability  $\varepsilon_*$ . The full conditional distributions of the Gibbs sampler for the factor analysis model with outliers are derived in Appendix C.

### 3.1 The marginal likelihood for factor analysis with outliers

Using the approach of Chib (1995) we will evaluate the marginal likelihood at the point

$$\hat{\theta} = (\hat{\theta}_0, \hat{\mathbf{A}} = \mathbf{0}, \hat{\mathbf{D}}_\vartheta = \mathbf{0}) \quad (20)$$

where  $\hat{\theta}_0 = (\hat{\mathbf{Z}}, \hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}})$  is the same point as for the factor analysis without outliers. Now we have the following factorization

$$p(\hat{\theta} | \mathbf{Y}) = p(\hat{\mathbf{D}}_\vartheta | \mathbf{Y}) \cdot p(\hat{\mathbf{A}} | \hat{\mathbf{D}}_\vartheta, \mathbf{Y}) \cdot p(\hat{\theta}_0 | \hat{\mathbf{A}}, \hat{\mathbf{D}}_\vartheta, \mathbf{Y}) \quad (21)$$

and

$$p(\hat{\theta}) = p(\hat{\theta}_0) \mathcal{N}[\mathbf{A}_*, \mathbf{I}_n \otimes \mathbf{P}_*] \prod_{i=1}^n \text{Ber}(\varepsilon_*). \quad (22)$$

1. Use the Gibbs run of  $J$  sample points of the ‘factor analysis with outliers’ program to calculate the ordinate:

$$\begin{aligned} p(\hat{\mathbf{D}}_\vartheta | \mathbf{Y}) &= \int \prod_{i=1}^n \text{Ber}(\varepsilon_{i**}) p(\theta | \mathbf{Y}) d\theta \\ &= \frac{1}{J} \sum_{j=1}^J \prod_{i=1}^n \text{Ber}(\varepsilon_{i**}^{(j)}) \end{aligned} \quad (23)$$

where the parameter of the  $i$ -th posterior density of the Bernoulli distribution is given by

$$\varepsilon_{i**}^{(j)} = \frac{c_i^{(j)}}{c_i^{(j)} + d_i^{(j)}},$$

with the ordinates

$$\begin{aligned} c_i^{(j)} &= \mathcal{N}_p[\mathbf{y}_i | \mathbf{a}^{(j)} \vartheta_i^{(j)} + \mathbf{\Lambda}_j \mathbf{z}_i^{(j)}], \\ d_i^{(j)} &= \mathcal{N}_p[\mathbf{y}_i | \mathbf{\Lambda}^{(j)} \mathbf{z}_i^{(j)}]. \end{aligned}$$

2. The ordinate for the second component can be obtained without a Gibbs sampling output by

$$p(\hat{\mathbf{A}} | \hat{\mathbf{D}}_\vartheta = \mathbf{0}, \mathbf{Y}) = \mathcal{N}_{p \times n}[\hat{\mathbf{A}} | \mathbf{A}_{**}, \mathbf{I}_n \otimes \mathbf{G}_{**}] = \prod_{i=1}^n \mathcal{N}[\hat{\mathbf{a}}_i | \mathbf{a}_{i**}, \mathbf{G}_{**}]. \quad (24)$$

It can be seen from the f.c.d. for  $\mathbf{A}$  that for  $\hat{\mathbf{D}}_\vartheta = \mathbf{0}$  the conditional distribution equals the prior distribution:

$$\mathbf{G}_{**} = \mathbf{I}_n \otimes \mathbf{P}_* \quad \text{and} \quad \text{vec } \mathbf{A}_{**} = \text{vec } \mathbf{A}_*.$$

3. Finally we can obtain the ordinate of the third factor  $p(\hat{\theta}_0 | \hat{\mathbf{A}}, \hat{\mathbf{D}}_{\vartheta}, \mathbf{Y})$  in (21) by the marginal likelihood output of the factor analysis program without outliers. This follows from the fact that the reduced Gibbs run at this step of the factor analysis program with outliers is equivalent to a factor analysis program without outliers.

The log marginal likelihood is now computed as

$$\log p(\mathbf{Y}) = \log p(\mathbf{Y} | \hat{\theta}) + \log p(\hat{\theta}) - \log p(\hat{\theta} | \mathbf{Y}) \quad (25)$$

where the likelihood part is given by

$$p(\mathbf{Y} | \hat{\theta}_1) = \mathcal{N}[\mathbf{Y} | \hat{\Lambda}\hat{\mathbf{Z}}, \mathbf{I}_n \otimes \hat{\Psi}], \quad (26)$$

which is the same value as for the factor analysis without outliers, since  $\hat{\mathbf{D}}_{\vartheta} = \mathbf{0}$ . Note that formula (23) is a simplification since the components for the location shifts  $\hat{\mathbf{A}}$  in (25) cancel out.

### 3.2 Model selection with Bayes factors

Posterior odds are used in Bayesian analysis to choose between two or more different models for the same data set. The basic formula for choosing between models  $M_1$  and  $M_2$  is

$$\text{posterior odds} = \text{Bayes factor} \cdot \text{prior odds}$$

or

$$\frac{p(M_1|\mathbf{Y})}{p(M_2|\mathbf{Y})} = B \cdot \frac{p(M_1)}{p(M_2)},$$

where  $p(M_1|\mathbf{Y})$  and  $p(M_2|\mathbf{Y})$  are the posterior probabilities for models  $M_1$  and  $M_2$ , respectively.  $p(M_1)$  and  $p(M_2)$  are the prior probabilities for models  $M_1$  and  $M_2$ , and, in the simplest case, they are set to be equal. Thus, in these cases the posterior odds are equal to the Bayes factor, which is defined as the ratio of marginal likelihoods

$$B = \frac{p(M_1|\mathbf{Y})}{p(M_2|\mathbf{Y})} = \frac{\int p(\mathbf{Y}, \theta_1) d\theta_1}{\int p(\mathbf{Y}, \theta_2) d\theta_2},$$

where  $\theta_1$  and  $\theta_2$  are the parameters for models  $M_1$  and  $M_2$ , respectively. If  $B > 1$  we choose model  $M_1$  and if  $B < 1$  we choose model  $M_2$ . Therefore

the model with the largest marginal likelihood will be chosen using simple Bayes factors. For example: The Bayes factor for testing the factor analysis model with outliers against no outliers is:

$$B = \frac{p(\mathbf{Y}|\text{outliers})}{p(\mathbf{Y}|\text{no outliers})}.$$

Clearly, this approach can be combined with testing the rank of the factor matrix by Bayes factors as in (11).

## 4 Example

The main purpose of factor analysis is the reduction of dimensionality. Therefore it will generally be difficult to come up with precise prior information. While classical factor analysis achieves identification of the model by non-stochastic restrictions, Bayesian factor analysis can be thought of as a stochastic weakening of these assumptions. Therefore we suggest using the eigenvalue decomposition of covariance matrices as a data-based prior distribution. This approach seems to be plausible since the goal is to explore the effects of outliers. The rationale of this specification is as follows: even if outliers are present, we expect the factors and the factor loadings to be in the "neighborhood" of the original model.

We use the language data in Fuller (1987, page 154) as an example for identifying outliers in a factor analysis. This data set consists of 100 observations with eight items: three items related to the essay, three items related to the language used, and two items related to the writing style.

We are interested to see if there are outliers in the 8-dimensional data set which might affect the factor analysis. The assignment of the prior parameters was based on the classical least squares approach to factor analysis in the following way (see Reyment and Joereskog (1996)):

1. We calculated the eigenvalue decomposition:

$$\begin{aligned} \mathbf{Y}\mathbf{Y}' &= \mathbf{U}\mathbf{\Gamma}\mathbf{U}', \\ \mathbf{U} &= (\mathbf{u}_1, \dots, \mathbf{u}_p), \\ \mathbf{\Gamma} &= \text{diag}(\gamma_1, \dots, \gamma_p). \end{aligned}$$

2. If we choose  $K = 2$  factors, then we use the first two eigenvectors of  $\mathbf{U}$  together with the largest two eigenvalues in  $\mathbf{\Gamma}$ :

$$\begin{aligned}\mathbf{U}_k &= (\mathbf{u}_1, \mathbf{u}_2), \\ \mathbf{\Gamma}_k &= \begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}.\end{aligned}$$

3. The prior parameters are now

$$\begin{aligned}\mathbf{\Lambda}_* &= \mathbf{Y}\mathbf{U}_k, \\ \mathbf{H}_* \otimes \mathbf{G}_* &= \mathbf{\Gamma}_k \otimes \mathbf{I}_p, \\ \mathbf{\Phi}_* &= \mathbf{I}_k, \\ \mathbf{\Psi}_* &= \text{diag}(\text{var}(\mathbf{y}_1), \dots, \text{var}(\mathbf{y}_p))/10.\end{aligned}$$

4. The prior distribution for the outlier parameters in model (14) consists of two parts. The first part is the set of parameters which is identical to the factor analysis model without outliers in section two. For the prior distribution of the location shifts we have assumed  $\mathbf{A}_* = \mathbf{0}$  and  $\text{var}(\mathbf{a}_i) = \mathbf{P}_* = \text{diag}(\text{var}(\mathbf{y}))$ . Finally, we assume that the residual variances of the factor model are about one tenth of the variances of the observed variables. The value of the prior information of the Wishart distribution is  $n_* = \nu_* = 1$ , i.e. 1/100 in terms of the sample size  $n = 100$ .

Convergence of the Gibbs sampler was achieved quite quickly for the present specification. The convergence was monitored by diagnostic measures proposed in the CODA package of Best et al. (1995) written in S-plus which uses the Gelman and Rubin (1992) and the Raftery and Lewis (1992) statistics. A good introduction to the theory and practice of MCMC modeling can be found in Gilks et al. (1990). Only the last 100 simulations of the MCMC sequence were used to calculate the mean and variances of the posterior distribution. Note that convergence could not be achieved for arbitrary prior distributions for factors and factor loadings. Quick and satisfactory results were only obtained for the proposed data-based prior distribution. This approach also guarantees an insensitive calculation of the marginal likelihood. Priors in the vicinity of the data-based prior will produce identical results for model selection via the marginal likelihood criterion and only little variation

j	1	2	3	4	5	6	7	8
$p_j$	0.739	0.793	0.842	0.879	0.914	0.948	0.977	1.000

Table 1: Cumulated percentages

for Bayes factors.

The eigenvalues of the data covariance matrix are

$$\text{diag}\mathbf{\Gamma} = (207.339, 15.079, 13.848, 10.185, 10.026, 9.400, 8.038, 6.572),$$

and the cumulated percentages are

$$p_j = \frac{\sum_{i=1}^j \gamma_i}{\sum_{i=1}^8 \gamma_i}, \quad j = 1, \dots, 8,$$

the results are given by Table 1. Table 2 lists the (ordinary) marginal likelihood and the posterior marginal likelihood (see Appendix D) for different number of factors. Two factors are chosen by the maximum marginal likelihood criterion for the model with and without outliers. The Bayes factor is clearly in favor of the factor analysis model with outliers.

Table 3 is the summary of the important result from our factor analysis model with outliers. The first column is the observation number and the second column is the probability of its being an outlier. The third to the tenth columns contain the outlier shifts and the standard deviations of the outliers in parentheses.

Table 3 shows the row estimates of the location shift matrix  $\mathbf{A}$  for which the posterior probability parameter  $\varepsilon_{i**}$  (the probability of being an outlier) is larger than 1/2. The prior probability that observation  $i$  is an outlier is assumed to be  $\varepsilon_{i*} = 0.1$ . The standard deviations of the location shifts are printed in parentheses. Those location shifts  $a_{ij}$  which are larger than the standard deviation are in bold font. It is interesting to note that all five outlier points have location shifts which are shifted by more than one standard deviation in exactly one of the eight variables. This shows that the grading process of the language papers was quite independent with respect to these eight judgment categories. No outlier point shows location shifts in *two* or more variables jointly. Note that the standard deviations of the

location shifts varies quite a lot across the outliers. There seems to be no obvious relation to the posterior probability of being an outlier. The size of the location shifts are not too large but make sense if they are compared to the original data.

Figure 1 shows the eight row vectors of the factor loadings matrix  $\mathbf{A}$  for the model with and without outliers. The eight row vectors of the factor loadings for the model without outliers are more spread out than the factor loadings for the factor model with the outliers removed.

Figure 2 shows the posterior probabilities of the location shifts in the factor analysis model of rank 2. The posterior means  $\varepsilon_{i**}$  are interpreted as posterior probabilities of being an outlier. This leads us to the following stochastic outlier analysis: While the posterior means for three observations are above 60%, two more observations just make it above the 50% line. Three (or almost four) additional observations are above 40% while for the other observations the  $\varepsilon_{**}$  are close to zero. We conclude that checking for outliers can be important for factor analysis with data sets which are exposed to possibly aberrant observations.

## 5 Conclusions

The paper shows that factor analysis can be combined quite successfully with a multivariate approach to outlier analysis. Under the assumption of a normal-Wishart distribution and a location-shift outlier model with Bernoulli-distributed indicator variables, all the full conditional distributions of the Gibbs sampler are given in closed form. The method is demonstrated with the Fuller (1987) language data and is tested for the presence of outliers. We have also shown that the problem of estimating the number of factors can be solved by using the concept of marginal likelihoods. The marginal likelihood can be computed as an additional step in the Gibbs sampling algorithm. With our example we could demonstrate that the marginal likelihood (and posterior marginal likelihood) criteria pick the same number of factors. The present approach uses a data-based prior distribution which yields a fast convergence of the Gibbs sampler. Furthermore Shera and Ibrahim (1998) have suggested using historical data for a prior distribution which could also be used in the context of outliers. Recently, Mori et al. (1998) developed a program package which performs sensitivity analysis for multivariate methods.

Applying the Fuller (1987) data set to a factor analysis with 2 factors identifies 8 to 10 observations as influential points which corresponds well with our analysis. Further research in this area will show if the present approach can be extended to more general distributional assumptions, alternative prior distributions or different outlier models.

## Appendix

Appendix A demonstrates that the multivariate regression model with outliers is a special case of the factor analysis with outlier model.

### A Multivariate Regression and Outliers

The multivariate regression model is given by

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{E}.$$

$(n \times p) \quad (n \times K) \quad (K \times p) \quad (n \times p)$

Let  $\mathbf{A}$  be a  $n \times p$  location shift parameter matrix and  $\mathbf{D}_\vartheta = \text{diag}(\vartheta_1, \dots, \vartheta_n)$  an  $n \times n$  indicator matrix for multivariate outliers. We can then formulate the model by assuming a normal distribution

$$\mathbf{Y} \sim \mathcal{N}_{n \times p}[\mathbf{X}\mathbf{B} + \mathbf{D}_\vartheta\mathbf{A}, \mathbf{\Psi} \otimes \mathbf{I}_n].$$

The prior information can be compactly formulated as

$$\begin{aligned} \mathbf{B} &\sim \mathcal{N}_{K \times p}[\mathbf{B}_*, \mathbf{G}_* \otimes \mathbf{H}_*], \\ \mathbf{\Psi} &\sim \mathcal{W}_p[\mathbf{\Psi}_*, n_*], \\ \mathbf{A} &\sim \mathcal{N}_{n \times p}[\mathbf{A}_*, \mathbf{P}_* \otimes \mathbf{I}_n], \\ \vartheta_i &\sim \text{Ber}[\varepsilon_{i*}], \quad i = 1, \dots, n, \end{aligned}$$

where  $\text{Ber}[\varepsilon_{i*}]$  denotes the Bernoulli distribution and  $\varepsilon_{i*}$  is the prior probability that observation  $i$  is an outlier.

The joint distribution of the data  $\mathbf{Y}$  and the parameter  $\theta = (\mathbf{B}, \mathbf{\Psi}, \mathbf{A}, \mathbf{D}_\vartheta)$  is

$$\begin{aligned} p(\mathbf{Y}, \theta) &= \mathcal{N}_{n \times p}[\mathbf{Y} \mid \mathbf{X}\mathbf{B} + \mathbf{D}_\vartheta\mathbf{A}, \mathbf{\Psi} \otimes \mathbf{I}_n] \cdot \mathcal{N}_{K \times p}[\mathbf{B} \mid \mathbf{B}_*, \mathbf{G}_* \otimes \mathbf{H}_*] \cdot \\ &\quad \cdot \mathcal{W}_p[\mathbf{\Psi}^{-1} \mid \mathbf{\Psi}_*, n_*] \cdot \mathcal{N}_{n \times p}[\mathbf{A} \mid \mathbf{A}_*, \mathbf{P}_* \otimes \mathbf{I}_n] \cdot \sum_{i=1}^n \text{Ber}(\vartheta_i \mid \varepsilon_{i*}). \end{aligned}$$

The full conditional distributions are

1. For the matrix regression coefficients  $\mathbf{B}$ :

$$p(\mathbf{B} \mid \mathbf{Y}, \theta^c) = \mathcal{N}_{n \times p}[\mathbf{B}_{**}, \mathbf{C}_{**}]$$

a multivariate normal distribution with the parameters

$$\begin{aligned} \mathbf{C}_{**}^{-1} &= \mathbf{G}_*^{-1} \otimes \mathbf{H}_*^{-1} + \boldsymbol{\Psi}^{-1} \otimes \mathbf{X}'\mathbf{X}, \\ \text{vec } \mathbf{B}_{**} &= \mathbf{C}_{**}^{-1} [\text{vec} (\mathbf{G}_*^{-1} \mathbf{B}_* \mathbf{H}_*^{-1} + \mathbf{X}'(\mathbf{Y} - \mathbf{D}_{\vartheta} \mathbf{A}) \boldsymbol{\Psi}^{-1})]. \end{aligned}$$

2. For the covariance matrix  $\boldsymbol{\Psi}$ :

$$p(\boldsymbol{\Psi}^{-1} \mid \mathbf{Y}, \theta^c) = \mathcal{W}_p[\boldsymbol{\Psi}_{**}, n_{**} = n_* + n]$$

a  $p$ -dimensional Wishart distribution with scale parameter

$$\boldsymbol{\Psi}_{**} = \boldsymbol{\Psi}_* + (\mathbf{Y} - \mathbf{X}\mathbf{B} - \mathbf{D}_{\vartheta} \mathbf{A})(\mathbf{Y} - \mathbf{X}\mathbf{B} - \mathbf{D}_{\vartheta} \mathbf{A})'.$$

3. For the level shift matrix  $\mathbf{A}$ :

$$p(\mathbf{A} \mid \mathbf{Y}, \theta^c) = \mathcal{N}_{n \times n}[\mathbf{A}_{**}, \mathbf{G}_{**}]$$

a multivariate normal distribution with the parameters

$$\begin{aligned} \mathbf{G}_{**}^{-1} &= \mathbf{P}_*^{-1} \otimes \mathbf{I}_n + \boldsymbol{\Psi}^{-1} \otimes \mathbf{D}'_{\vartheta} \mathbf{D}_{\vartheta}, \\ \text{vec } \mathbf{A}_{**} &= \mathbf{G}_{**}^{-1} [\text{vec} (\mathbf{A}_* \mathbf{P}_*^{-1} + \mathbf{D}_{\vartheta} (\mathbf{Y} - \mathbf{X}\mathbf{B}) \boldsymbol{\Psi}^{-1})]. \end{aligned}$$

For each observation the posterior mean can be calculated by breaking up the system estimate as

$$\begin{aligned} \mathbf{G}_{**i}^{-1} &= \mathbf{P}_*^{-1} + \vartheta_i^2 \boldsymbol{\Psi}^{-1}, \quad i = 1, \dots, n, \\ \mathbf{a}_{**i} &= \mathbf{G}_{**i}^{-1} [\mathbf{P}_*^{-1} \mathbf{a}_{*i} + \vartheta_i \boldsymbol{\Psi}^{-1} (\mathbf{y}_i - \mathbf{B}\mathbf{x}_i)]. \end{aligned}$$

4. For the indicator variables  $\vartheta_i$ :

$$Pr(\vartheta_i \mid \mathbf{Y}, \theta^c) = Ber[\varepsilon_{i**} = \frac{c_i}{c_i + d_i}], \quad i = 1, \dots, n,$$

a Bernoulli distribution with the components obtained via Bayes theorem, i.e.,

$$\begin{aligned} c_i &= \mathcal{N}_p[\mathbf{y}_i \mid \mathbf{x}_i \mathbf{B} + \mathbf{a}_i, \boldsymbol{\Psi}] \cdot \varepsilon_{i*}, \\ d_i &= \mathcal{N}_p[\mathbf{y}_i \mid \mathbf{x}_i \mathbf{B}, \boldsymbol{\Psi}] \cdot (1 - \varepsilon_{i*}), \quad i = 1, \dots, n, \end{aligned}$$

where  $\mathbf{x}_i$  is the  $i$ -th row of  $\mathbf{X}$  and  $\mathbf{a}_i$  is the  $i$ -th row of  $\mathbf{A}$ .

## B The computation of the marginal likelihood

This appendix calculates the marginal likelihood of the factor analysis model (see section 2.1) using the method of Chib (1995).

1. Use the Gibbs run of length  $J$  of the estimation procedure to estimate the ordinate at the location  $\hat{\mathbf{Z}}$ :

$$\begin{aligned} p(\hat{\mathbf{Z}} | \mathbf{Y}) &= \int \mathcal{N}[\hat{\mathbf{Z}} | \mathbf{z}_{**}, \mathbf{I}_n \otimes \mathbf{D}_Z] \times p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Phi}^{-1} | \mathbf{Y}) d\boldsymbol{\Lambda} d\boldsymbol{\Psi}^{-1} d\boldsymbol{\Phi}^{-1} \\ &= \frac{1}{J} \sum_{j=1}^J \mathcal{N}[\hat{\mathbf{Z}} | \mathbf{z}_{**}^{(j)}, \mathbf{I}_n \otimes \mathbf{D}_Z^{(j)}] = \frac{1}{J} \sum_{j=1}^J \prod_{i=1}^n \mathcal{N}[\hat{z}_i | \mathbf{z}_{**i}^{(j)}, \mathbf{D}_Z^{(j)}], \end{aligned} \quad (27)$$

with  $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$  and the following moments of the multivariate normal distribution

$$\begin{aligned} \mathbf{z}_{**}^{(j)} &= (\boldsymbol{\Phi}_j^{-1} + \boldsymbol{\Lambda}_j' \boldsymbol{\Psi}_j^{-1} \boldsymbol{\Lambda}_j)^{-1} \boldsymbol{\Lambda}_j' \boldsymbol{\Psi}_j^{-1} \mathbf{Y}, \\ \mathbf{D}_Z^{(j)} &= (\boldsymbol{\Phi}_j^{-1} + \boldsymbol{\Lambda}_j' \boldsymbol{\Psi}_j^{-1} \boldsymbol{\Lambda}_j)^{-1}. \end{aligned}$$

2. Reduce the Gibbs run to estimate the second component of the posterior ordinate

$$\begin{aligned} p(\hat{\boldsymbol{\Lambda}} | \hat{\mathbf{Z}}, \mathbf{Y}) &= \int \mathcal{N}[\hat{\boldsymbol{\Lambda}} | \boldsymbol{\Lambda}_{**}, \mathbf{C}_{**}] \times p(\boldsymbol{\Psi}^{-1}, \boldsymbol{\Phi}^{-1} | \mathbf{Y}) d\boldsymbol{\Psi}^{-1} d\boldsymbol{\Phi}^{-1} \\ &= \frac{1}{J} \sum_{j=1}^J \mathcal{N}[\hat{\boldsymbol{\Lambda}} | \boldsymbol{\Lambda}_{**}^{(j)}, \mathbf{C}_{**}^{(j)}] \end{aligned} \quad (28)$$

with the reduced parameters

$$\boldsymbol{\Lambda}_{**}^{(j)} = \mathbf{C}_{**}^{(j)} [\text{vec} (\mathbf{G}_*^{-1} \boldsymbol{\Lambda}_* \mathbf{H}_*^{-1} + \boldsymbol{\Psi}_j^{-1} \mathbf{Y} \hat{\mathbf{Z}})]$$

and

$$\mathbf{C}_{**}^{(j)} = (\mathbf{H}_*^{-1} \otimes \mathbf{G}_*^{-1} + \hat{\mathbf{Z}}' \hat{\mathbf{Z}} \otimes \boldsymbol{\Psi}_j^{-1})^{-1}, \quad j = 1, \dots, J.$$

Note that this implies that the remaining two conditional distributions have to be estimated by

$$\boldsymbol{\Psi}_{**} = \boldsymbol{\Psi}_* + (\mathbf{Y} - \hat{\boldsymbol{\Lambda}} \hat{\mathbf{Z}})(\mathbf{Y} - \hat{\boldsymbol{\Lambda}} \hat{\mathbf{Z}})' \quad (29)$$

and

$$\Phi_{**} = \Phi_* + \hat{\mathbf{Z}}\hat{\mathbf{Z}}'. \quad (30)$$

3. The third component of the posterior ordinate can be estimated without a Gibbs run because

$$\begin{aligned} p(\hat{\Psi}^{-1} | \hat{\Lambda}, \hat{\mathbf{Z}}, \mathbf{Y}) &= \int \mathcal{W}_p[\hat{\Psi}^{-1} | \hat{\Lambda}, \hat{\mathbf{Z}}, \mathbf{Y}] d\Phi^{-1} \\ &= \mathcal{W}_p[\hat{\Psi}^{-1} | \Psi_{**}, n_{**}], \end{aligned} \quad (31)$$

which does not depend on the remaining  $\Phi$  parameters and with  $\Psi_{**}$  given as in (29).

4. The last component of the posterior ordinate can be also calculated without a Gibbs run:

$$p(\hat{\Phi}^{-1} | \hat{\Psi}^{-1}, \hat{\Lambda}, \hat{\mathbf{Z}}, \mathbf{Y}) = \mathcal{W}[\hat{\Phi}^{-1} | \Phi_{**}, \nu_{**}], \quad (32)$$

where  $\Phi_{**}$  is given in (30).

## C The Gibbs sampler for the factor analysis with outliers

The joint distribution of the parameters  $\theta = (\Lambda, \mathbf{Z}, \Psi, \Phi, \mathbf{A}, \vartheta)$  and the data  $\mathbf{Y}$  is found from the distributions in (18) and (20):

$$\begin{aligned} p(\mathbf{Y}, \theta) &= \mathcal{N}_{p \times n}[\mathbf{Y} | \mathbf{A}\mathbf{D}\vartheta + \Lambda\mathbf{Z}, \mathbf{I}_n \otimes \Psi] \times \mathcal{W}_p[\Psi^{-1} | \Psi_*, n_*] \\ &\quad \times \mathcal{N}_{K \times n}[\mathbf{Z} | \mathbf{0}, \mathbf{I}_n \otimes \Phi] \times \mathcal{W}_K[\Phi^{-1} | \Phi_*, \nu_*] \\ &\quad \times \mathcal{N}_{p \times K}[\Lambda | \Lambda_*, \mathbf{H}_* \otimes \mathbf{G}_*] \times \mathcal{N}_{p \times K}[\mathbf{A} | \mathbf{A}_*, \mathbf{I}_K \otimes \mathbf{P}_*] \\ &\quad \times \prod_{i=1}^n \text{Ber}[\vartheta_i, \varepsilon_*]. \end{aligned}$$

The full conditional distributions are:

1. For the latent factors  $\mathbf{Z}$ :

$$\begin{aligned}
p(\mathbf{Z} \mid \mathbf{Y}, \theta^c) &\propto \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta - \mathbf{\Lambda}\mathbf{Z})'\mathbf{\Psi}^{-1}(\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta - \mathbf{\Lambda}\mathbf{Z})\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\mathbf{Z}'\mathbf{\Phi}^{-1}\mathbf{Z}\right\} \\
&= \mathcal{N}_{K \times n}[\mathbf{Z}_{**}, \mathbf{I}_n \otimes \mathbf{D}_Z].
\end{aligned} \tag{33}$$

Because  $\text{vec } \mathbf{\Lambda}\mathbf{Z} = (\mathbf{I}_n \otimes \mathbf{\Lambda}) \text{vec } \mathbf{Z}$ , we find for the parameters

$$\mathbf{D}_Z^{-1} = \mathbf{\Phi}^{-1} + \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda},$$

and

$$\mathbf{Z}_{**} = \mathbf{D}_Z \mathbf{\Lambda}' \mathbf{\Psi}^{-1} (\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta).$$

2. For the matrix of factor loadings  $\mathbf{\Lambda}$ :

$$\begin{aligned}
p(\mathbf{\Lambda} \mid \mathbf{Y}, \theta^c) &\propto \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta - \mathbf{\Lambda}\mathbf{Z})'\mathbf{\Psi}^{-1}(\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta - \mathbf{\Lambda}\mathbf{Z})\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\text{tr}\mathbf{H}_*^{-1}(\mathbf{\Lambda} - \mathbf{\Lambda}_*)'\mathbf{G}_*^{-1}(\mathbf{\Lambda} - \mathbf{\Lambda}_*)\right\} \\
&= \mathcal{N}_{p \times K}[\mathbf{\Lambda}_{**}, \mathbf{C}_{**}]
\end{aligned} \tag{34}$$

with the parameters

$$\mathbf{C}_{**}^{-1} = \mathbf{H}_*^{-1} \otimes \mathbf{G}_*^{-1} + \mathbf{Z}\mathbf{Z}' \otimes \mathbf{\Psi}^{-1}$$

and

$$\text{vec } \mathbf{\Lambda}_{**} = \mathbf{C}_{**}[\text{vec } (\mathbf{G}_*^{-1}\mathbf{\Lambda}_*\mathbf{H}_*^{-1} + \mathbf{\Psi}^{-1}(\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta)\mathbf{Z})]$$

since also  $\text{vec } \mathbf{\Lambda}\mathbf{Z} = (\mathbf{Z}' \otimes \mathbf{I}_p) \text{vec } \mathbf{\Lambda}$  and

$$\text{vec } \mathbf{\Lambda}_{**} = \mathbf{C}_{**}[(\mathbf{H}_*^{-1} \otimes \mathbf{G}_*^{-1}) \text{vec } \mathbf{\Lambda}_* + (\mathbf{Z} \otimes \mathbf{\Psi}^{-1})\text{vec } (\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta)].$$

3. For the covariance matrix  $\mathbf{\Psi}$ :

$$p(\mathbf{\Psi}^{-1} \mid \mathbf{Y}, \theta^c) = \mathcal{W}_p[\mathbf{\Psi}_{**}, n_{**} = n_* + n] \tag{35}$$

with

$$\mathbf{\Psi}_{**} = \mathbf{\Psi}_* + (\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta - \mathbf{\Lambda}\mathbf{Z})(\mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta - \mathbf{\Lambda}\mathbf{Z})'$$

because the residual matrix is  $\mathbf{E} = \mathbf{Y} - \mathbf{A}\mathbf{D}_\vartheta - \mathbf{\Lambda}\mathbf{Z}$ , and the f.c.d. is proportional to

$$p(\mathbf{\Psi}^{-1} \mid \mathbf{Y}, \theta^c) \propto |\mathbf{\Psi}^{-1}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}tr\mathbf{E}'\mathbf{\Psi}^{-1}\mathbf{E}\right\} \\ |\mathbf{\Psi}^{-1}|^{\frac{n_*-p-1}{2}} \exp\left\{-\frac{1}{2}tr\mathbf{\Psi}^{-1}\mathbf{\Psi}_*\right\}.$$

4. For the covariance matrix  $\mathbf{\Phi}$ :

$$p(\mathbf{\Phi}^{-1} \mid \mathbf{Y}, \theta^c) \propto |\mathbf{\Phi}^{-1}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}tr\mathbf{Z}'\mathbf{\Phi}^{-1}\mathbf{Z}\right\} \quad (36)$$

$$|\mathbf{\Phi}^{-1}|^{\frac{\nu_*-K-1}{2}} \exp\left\{-\frac{1}{2}tr\mathbf{\Phi}^{-1}\mathbf{\Phi}_*\right\} \quad (37)$$

$$= \mathcal{W}_K[\mathbf{\Phi}_{**}, \nu_{**} = \nu_* + n] \quad (38)$$

we find as f.c.d. a Wishart distribution with scale matrix

$$\mathbf{\Phi}_{**} = \mathbf{\Phi}_* + \mathbf{Z}\mathbf{Z}'.$$

5. For the level shift matrix  $\mathbf{A}$ :

$$p(\mathbf{A} \mid \mathbf{Y}, \theta^c) \propto \exp\left\{-\frac{1}{2}tr\mathbf{E}'\mathbf{\Psi}^{-1}\mathbf{E}\right\} \quad (39)$$

$$\times \exp\left\{-\frac{1}{2}tr(\mathbf{A} - \mathbf{A}_*)'\mathbf{P}_*^{-1}(\mathbf{A} - \mathbf{A}_*)\right\} \quad (40)$$

$$= \mathcal{N}_{p \times n}[\mathbf{A}_{**}, \tilde{\mathbf{G}}_{**}]. \quad (41)$$

Because of the vectorization

$$\text{vec } \mathbf{A}\mathbf{D}_\vartheta = (\mathbf{D}'_\vartheta \otimes \mathbf{I}_n) \text{vec } \mathbf{A},$$

we have for the parameters of the multivariate normal distribution

$$\tilde{\mathbf{G}}_{**}^{-1} = \mathbf{I}_n \otimes \mathbf{P}_*^{-1} + \mathbf{D}_\vartheta \mathbf{D}'_\vartheta \otimes \mathbf{\Psi}^{-1}, \\ \text{vec } \mathbf{A}_{**} = \tilde{\mathbf{G}}_{**}[(\mathbf{I}_n \otimes \mathbf{P}_*^{-1}) \text{vec } \mathbf{A}_* + (\mathbf{D}_\vartheta \otimes \mathbf{\Psi}^{-1}) \text{vec } (\mathbf{Y} - \mathbf{\Lambda}\mathbf{Z})] \\ = \tilde{\mathbf{G}}_{**}[\text{vec } (\mathbf{P}_*^{-1}\mathbf{A}_* + \mathbf{\Psi}^{-1}(\mathbf{Y} - \mathbf{\Lambda}\mathbf{Z})\mathbf{D}'_\vartheta)].$$

We can calculate the posterior mean for each observation as follows

$$\tilde{\mathbf{G}}_{**}^{-1} = \mathbf{P}_*^{-1} + \vartheta_i^2 \mathbf{\Psi}^{-1}, \\ \mathbf{a}_{**i} = \tilde{\mathbf{G}}_{**}[\mathbf{P}_*^{-1} + \vartheta_i \mathbf{\Psi}^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{z}_i)].$$

6. For the indicator variables  $\vartheta_i$  the f.c.d. is

$$p(\vartheta_i | \mathbf{Y}, \theta^c) = \text{Ber} \left[ \varepsilon_{i**} = \frac{c_i}{c_i + d_i} \right], \quad i = 1, \dots, n, \quad (42)$$

we obtain again a Bernoulli distribution for each observation by a straightforward application of the Bayes theorem with the components

$$\begin{aligned} c_i &= \text{Pr}(\mathbf{y}_i | \vartheta_i = 1, \mathbf{Y}, \theta^c) \varepsilon_* \\ &= \mathcal{N}_p[\mathbf{y}_i | \mathbf{a}_i + \mathbf{\Lambda} \mathbf{z}_i, \mathbf{\Psi}] \varepsilon_* \end{aligned}$$

and

$$\begin{aligned} d_i &= \text{Pr}(\mathbf{y}_i | \vartheta_i = 0, \mathbf{Y}, \theta^c) (1 - \varepsilon_*) \\ &= \mathcal{N}_p[\mathbf{y}_i | \mathbf{\Lambda} \mathbf{z}_i, \mathbf{\Psi}] (1 - \varepsilon_*), \quad i = 1, \dots, n, \end{aligned}$$

where  $\mathbf{z}_i$  is the  $i$ -th column of  $\mathbf{Z}$ ,  $\vartheta_i$  is the  $i$ -th column of  $\mathbf{D}_\vartheta$ , and  $\mathcal{N}_p$  is the  $p$ -dimensional normal density function.

The posterior probability that observation  $i$  is an outlier is estimated from the MCMC output for the Bernoulli parameter  $\varepsilon_{i**}$ :

$$\bar{\varepsilon}_{i**} = \frac{1}{J} \sum_{j=1}^J \varepsilon_{i**}^{(j)}.$$

## D The posterior marginal likelihood

Since the computation of the marginal likelihoods needs considerable computation time, we propose the calculation of the posterior marginal likelihood (see Aitkin 1991) from the MCMC output for large data sets.

### 1. The factor analysis model:

The posterior marginal likelihood (*PoML*) of the factor analysis model of rank  $K$  can be estimated from the MCMC output of size  $R$  in following way:

$$\begin{aligned} \text{PoML} &= \frac{1}{R} \sum_{j=1}^R \mathcal{N}[\mathbf{Y} | \mathbf{\Lambda}_{(j)} \mathbf{Z}_{(j)}, \mathbf{I}_n \otimes \mathbf{\Psi}_{(j)}] \cdot \mathcal{N}[\mathbf{Z}_{(j)} | \mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Phi}_{(j)}] \cdot \\ &\quad \mathcal{N}[\mathbf{\Lambda}_{(j)} | \mathbf{\Lambda}_*, \mathbf{H}_* \otimes \mathbf{G}_*] \cdot \mathcal{W}[\mathbf{\Psi}_{(j)}^{-1} | \mathbf{\Psi}_*, n_*] \cdot \mathcal{W}[\mathbf{\Phi}_{(j)}^{-1} | \mathbf{\Phi}_*, \nu_*]. \end{aligned}$$

## 2. The factor analysis model with outliers:

The *PoML* for the factor analysis model with outliers is given by

$$\begin{aligned}
 PoML &= \frac{1}{R} \sum_{j=1}^R \mathcal{N}[\mathbf{Y} | \mathbf{\Lambda}_{(j)} \mathbf{Z}_{(j)} + \mathbf{A}_{(j)} \mathbf{D}_{\theta}^{(j)}, \mathbf{I}_n \otimes \mathbf{\Psi}_{(j)}] \cdot \mathcal{N}[\mathbf{Z}_{(j)} | \mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Phi}_{(j)}] \cdot \\
 &\quad \cdot \mathcal{N}[\mathbf{\Lambda}_{(j)} | \mathbf{\Lambda}_*, \mathbf{H}_* \otimes \mathbf{G}_*] \cdot \mathcal{N}[\mathbf{A}_{(j)} | \mathbf{A}_*, \mathbf{I}_n \otimes \mathbf{P}_*] \cdot \\
 &\quad \cdot \mathcal{W}[\mathbf{\Psi}_{(j)}^{-1} | \mathbf{\Psi}_*, n_*] \cdot \mathcal{W}[\mathbf{\Phi}_{(j)}^{-1} | \mathbf{\Phi}_*, \nu_*] \cdot \prod_{i=1}^n Ber(\theta_i^{(j)} | \varepsilon_{i*}).
 \end{aligned}$$

We have shown in several simulation runs that the *PoML* criterion chooses always the same model as the ordinary marginal likelihood criterion. This result can be expected since we use a data-based prior. Therefore we can recommend the *PoML* criterion for model selection for large data sets, since it is faster to compute.

## 5 References

- Aitkin, M. (1991): Posterior Bayes Factors, *J. Royal Statistical Society*, 53, 111-142.
- Akaike, H. (1987): Factor analysis of AIC, *Psychometrika* 52, 317-332.
- Barnett, V. and Lewis, T. (1984): *Outliers in Statistical Data*, Wiley, Chichester.
- Best, N.G.; Cowles, M.K. and Vines, K. (1995): CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, Version 3.0, Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.
- Chib, S. (1995): Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313-1321.
- Fuller, W.A. (1987): *Measurement Error Models*, John Wiley & Sons, NY.
- Gelfand, A.E. and Smith, A.F.M. (1990): Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.
- Gelman, A. and Rubin, D.B. (1992): Inference from iterative simulation using multiple sequences (with discussion), *Stat. Science*, 7, 457-511.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1990): *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.
- Kitagawa, G. and Akaike, H. (1992): A Quasi Bayesian Approach to Outlier Detection, *Ann. Inst. Stat. Mathematics*, 34, 95-104.
- Magnus, J.R. and Neudecker, H. (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, Chichester.
- Martin, J.L. and McDonald, R.P. (1975): Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika*, 40, 505-517.
- Mori, M. Watadani, S., Tanrumi, T. and Tanaka, Y. (1998): Development of Statistical Software SAMMIF for Sensitivity Analysis in Multivariate Method, *COMPSTAT*, 395-400, Physica-Verlag .
- Pettit, L.I. and Smith, A.F.M. (1985): Outliers and Influential Observations in Linear Models, in J.M. Bernardo et al. (eds.) *Bayesian Statistics*, 2, 473-

474, Elsevier Amsterdam

Polasek, W. (1999): Gibbs sampling in B-VAR models with latent variables. in: Innovations in multivariate statistical analysis-A Festschrift for Heinz Neudecker, R. J. Heijmans, D. S. G. Pollock and A. Satorra, (eds), *Kluwer Academic Publishers, Dordrecht*, 137-156.

Polasek, W. (ed.) (1998): The BASEL package, ISO-WWZ, University of Basel, mimeo. <ftp://iso.iso.unibas.ch/pub/basel>

Press, S.J. and Shigemasu, K. (1989): Bayesian inference in factor analysis. In Contributions to Probability and Statistics: *Essays in honor of Ingram Olkin*; L.J. Gleser, M.D. Perlman, S.J. Press and A.R. Sampson (eds.), 271-287. New York: Springer Verlag.

Raftery, A.E. and Lewis, S.M. (1992): One long run with diagnostics: implementation strategies for Markov chain Monte Carlo, *Stat. Science*, 7, 493-497.

Reyment, R. and Joereskog, K.G. (1996): *Applied Factor Analysis in the Natural Sciences*, Cambridge University Press.

Rocke, D.M. and Woodruff, D.L. (1996): Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association*, 91/435, 1047-1061.

Shera D.M. and Ibrahim, J.G. (1998): *Prior Elicitation and Computation for Bayesian Factor Analysis*, mimeo, Harvard School of Public Health.

Verdinelli, I. and Wasserman, L. (1991): Bayesian analysis of outlier problems using the Gibbs sampler, *Statistics and Computing 1991-1*, 105-117.

$K$	marginal likelihood		posterior marginal likelihood	
	without outliers	with outliers	without outliers	with outliers
1	-403.5833	-383.1141	-741.2623	-668.5273
2	-363.5028*	-354.0389*	-701.9056*	-649.7650*
3	-449.0960	-363.7826	-753.5612	-678.6722
4	-530.0755	-393.5813	-828.7373	-701.5263
5	-573.8466	-453.0068	-863.5635	-711.6621

Table 2: The log marginal and posterior marginal likelihoods for the factor analysis model of the language data in Fuller (1987).

( \* maximum marginal likelihood )

Obs.	Prob.	Location Shift (std)			
		Developed	Logical	Irritating	Intelligent
1	0.7205	<b>-1.1350(1.0561)</b>	-0.0456(1.1870)	-0.0159(1.2950)	-0.2289(0.9911)
11	0.5527	0.0901(1.0091)	0.2068(1.300)	-0.0813(0.9647)	-0.0646(1.4020)
37	0.7176	-0.3747(1.2241)	0.0777(1.2141)	-0.2365(1.2600)	<b>-0.1828(0.0330)</b>
80	0.5209	0.1028(1.1800)	0.0641(0.9988)	<b>-0.1777(1.1400)</b>	-0.1433(0.8869)
85	0.6221	0.0164(1.0520)	-0.2075(1.2250)	0.2944(1.1372)	<b>0.5053(0.2030)</b>
Obs.	Prob.	Location Shift (std)			
		Understand	Appropriate	Acceptable	Careful
1	0.7205	-0.4626(1.3260)	0.2021(1.0720)	-0.1748(1.2140)	0.2380(1.2891)
11	0.5527	-0.1661(1.0671)	0.0721(1.3611)	<b>-0.0513(0.0358)</b>	0.0015(1.2740)
37	0.7176	-0.0440(1.2831)	0.0862(1.1190)	-0.2429(1.0380)	<b>-0.1124(0.0850)</b>
80	0.5209	0.0189(1.0540)	0.0638(0.8845)	-0.1217(1.1905)	-0.04914(0.8681)
85	0.6221	0.0003(1.0360)	-0.0066(1.1590)	-0.2909(1.1520)	-0.2381(1.3810)

Table 3: The probability of being an outlier and the posterior mean of location shifts and standard deviations in the factor analysis ( $K = 2$ ) with outliers model of the language data in Fuller (1987). Posterior means larger than posterior standard deviations are bold face.

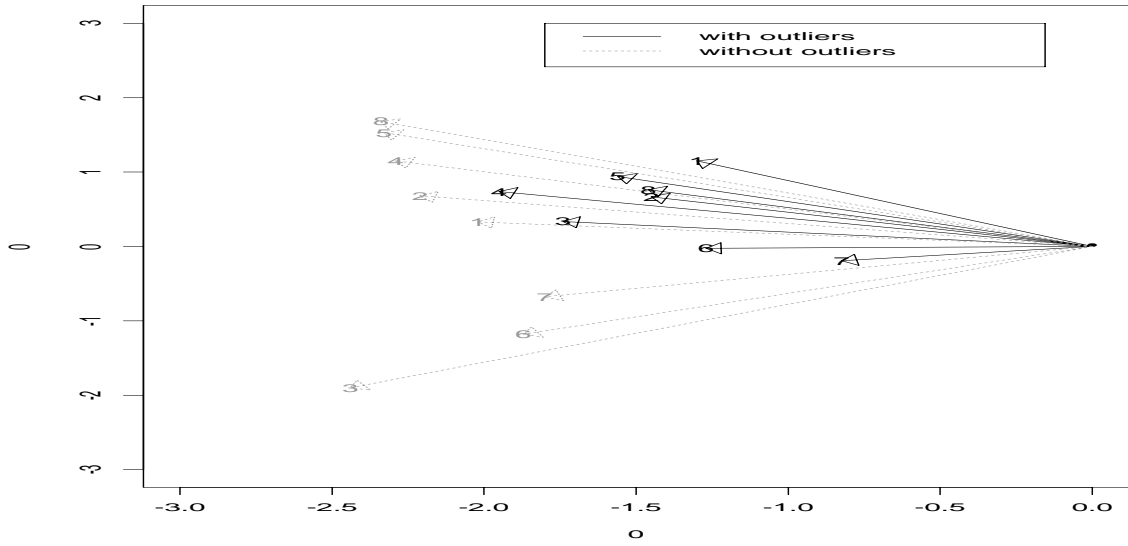


Figure 1: Language data: Factor loadings for the model with and without outliers.

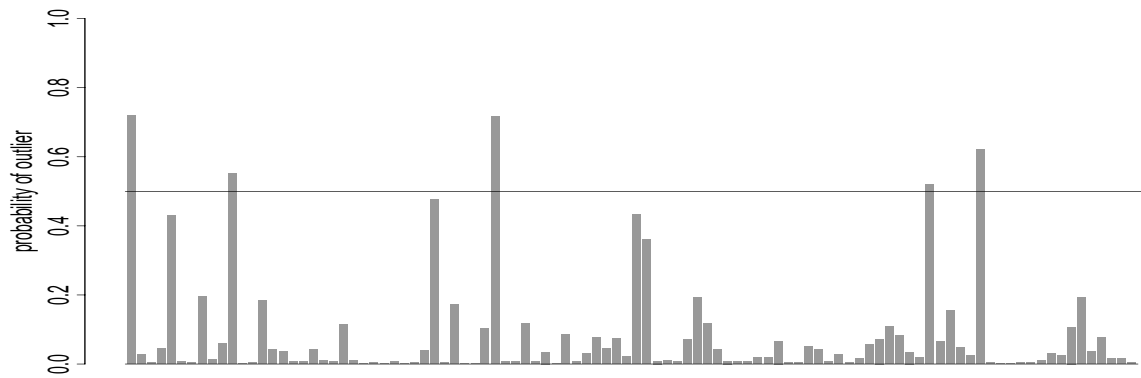


Figure 2: The probability of being an outlier for the language data in Fuller (1987) for the factor analysis ( $K = 2$ ) with outliers model.