THE · UNIVERSITY · OF · CALIFORNIA

LET THERE BE LIGHT

· 1868 ·

UNIVERSITY OF CALIFORNIA

RIVERSIDE

Correlated Bayesian Factor Analysis

A dissertation submitted in partial satisfaction

of the the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Daniel Bryant Rowe

December, 1998

Dissertation Committee:

Professor S. James Press, Chairperson

Professor Barry C. Arnold

Professor Bir Bhanu

The dissertation of Daniel Bryant Rowe is approved:

Committee Chairman

University of California, Riverside

To My Wife Gretchen,

My Mother Elizabeth,

My Brother Stephen.

Acknowledgments

There were many people that have contributed to my education and I wish to thank all of you.

I would like to thank the students, staff, and faculty at the University of California, Riverside. I would like to specifically thank Linda Penas and D.V. Gokhale for our Friday afternoon meetings; my office mate Mark Lehr for the great conversations we have had; and my friends, Carlos Lopez and Grinnell Jones.

I wish to thank my committee members Dr. Barry C. Arnold and Dr. Bir Bhanu. I wish to give special thanks to my advisor, Dr. S. James Press for his advice, support, and encouragement.

Daniel B. Rowe

ABSTRACT OF THE DISSERTATION

Correlated Bayesian Factor Analysis

by

Daniel Bryant Rowe

Doctor of Philosophy, Graduate program in Applied Statistics

University of California, Riverside, December 1998

Professor, S. James Press, Chairman

Factor analysis is a method in multivariate statistical analysis that can help scientists determine which variables to study in a field and their relationships. We extend the Bayesian approach to factor analysis developed in 1989 by Press and Shigemasu (henceforth PS89) and revised in 1997 to model correlated observation vectors, factor score vectors, and factor loadings. Further, we place a prior distribution on the number of factors and obtain posterior estimates.

Hitherto, factor analysis has only considered independent observation vectors. Quite often as evidenced by the large literature in time series and multivariate analysis, observation vectors are correlated. If correlation across observation vectors exists and this correlation is not taken into account, then the covariance matrix that is factor analyzed is improperly estimated.

Due to the number of parameters, a multivariate rejection sampling technique, and the computation involved for carrying out Gibbs sampling, we consider some covariance/correlation simplifications.

In appendices, we outline parameter estimation methods, hyperparameter assessment, correlation structure determination, and development for a general unknown mean for the observations with a prior distribution placed on it and estimated a posteriori with the other parameters.

Throughout, we will assume natural conjugate prior distributions for the unknown but fixed parameters, their posterior distribution will be formed, conditional posterior distributions will be found, and marginal posterior estimators will be calculated using Gibbs sampling.

The advantage of this extension and the Press and Shigemasu Bayesian approach to factor analysis is that they allow prior information that is available to formally be brought to bear, and at the same time, by incorporating proper prior information eliminate the indeterminencies or identification-of-parameters problem of classical factor analysis. There is no need for rotation of the factor loadings, the factor loadings are automatically found.

In the simulation example where the observations were correlated, we found that the correlated Bayesian factor analysis model determined the number of factors correctly and with a single parameter ρ performed better in estimating the parameters as evidenced by several performance measures.

In an example involving plankton, we found five underlying factors. The factors correspond to the four climate zones Tropical, Subtropical, Polar, Subpolar, and the Gyre margin assemblege region which is what the original authors found in their factor analysis.

Contents

List of Tables

1 Introduction and Methodology

1.1 Introduction

In a discipline, scientists are attempting to discover the relationships among variables. Knowledge in a discipline proceeds by determining which variables are related and to what extent. Factor analysis is a method in multivariate statistical analysis that can help determine which variables to study and their relationships. It uses the correlations or covariances between a set of observed variables to describe them in terms of a smaller set of unobservables. Factor analysis takes advantage of the relationships or correlation pattern within a set of vector valued observable random variables in order to describe them in terms of another set of vector valued unobservable or latent random variables called factors that are of lower dimension.

There are two main reasons why one would perform a factor analysis. The first is to explain the observed relationship among a set of observed variables in terms of a smaller number of hypothetical variables or latent factors which underlie the observations. This smaller number of variables can be used to find a meaningful structure in the observed variables. This structure will aid in the interpretation and explanation of the process that the has generated the observations.

The second reason one would carry out a factor analysis is for data reduction. Since we represent our observed variables in terms of a smaller number of unobserved or latent variables we reduce or minimize the number of variables in our analysis and hence reduce storage requirements. By having a smaller number of factors (vectors of smaller dimension) to work with that capture the essence of the observed variables, we only are required to store this smaller number of factors. We can also use the smaller number of factors for further analysis to reduce computational requirements.

Since factor analysis is a method that takes advantage of relationships among variables, we should model and take advantage of all possible relationships. Factor analysis has hitherto only taken advantage of the within observation vector relationships. We will take advantage of the possible between observation vector relationships as well by modeling a full covariance matrix for the observation vectors. We will also take advantage of the possible relationships between the unobservable factors by modeling a full covariance matrix for the factor vectors.

1.2 The Need for Correlation

The factor analysis model is based on the covariance matrix within the observation vectors. Previously, factor analysis models have assumed independence among the observation vectors. The observation vectors are often correlated as evidenced by the large literature in time series and multivariate analysis. If this correlation is not taken into account, then estimates of factor structure are relatively inefficient.

1.3 The Need for the Bayesian Approach

The use of the Bayesian approach to factor analysis bears directly on the problem of inestimability of parameters. In the classical factor analysis model, the parameters are inherently indeterminate. The parameters cannot be uniquely determined from the likelihood alone. When differentiating the log likelihood with respect to the unknown parameters and setting the resulting equations to zero, the system of equations for the parameters is not uniquely solvable in terms of the observations.

Early factor analysis methods as we will see, imposed constraints on the parameters in order to obtain determinate estimates within an orthogonal rotation. In PS89 the authors showed that these constraints are not necessary with the Bayesian model. Bayesian methods with the incorporation of available proper prior information eliminates the problem of indeterminacies.

When we attempt to account for correlation across the observation vectors, many new parameters are introduced and the the problem of indeterminacies is greatly increased. To remedy this, we use proper prior information and take advantage of simplifications of the covariance structures to reduce the number of distinct covariance terms.

2 Literature Review

This section is a review of the factor analysis model and some methods for estimating its parameters.

2.1 Factor Analysis Model

To explicate the factor analysis model we adopt a context in which each of many subjects is asked a battery of questions. Let x_i denote the p-vector of responses of subject i to p questions; $i = 1, \ldots, N$.

The model is

$$\underset{(p\times 1)}{(x_i|\mu,\Lambda,f_i,m)} = \underset{(p\times 1)}{\mu} + \underset{(p\times m)}{\Lambda}\;\underset{(m\times 1)}{f_i} + \underset{(p\times 1)}{\epsilon_i}\quad, \tag{2.1.1}$$

where

$\mu =$ a p-dimensional unobserved population mean vector,

$\Lambda =$ the $p \times m$ matrix of unobserved constants called the factor loading matrix $\Lambda = (\lambda_1', \ldots, \lambda_p')'$,

$f_i =$ an m-dimensional vector of unobservable "common" factor scores for the i^{th} subject, and

$\epsilon_i =$ a p-dimensional vector of "specific" errors or disturbance terms of the i^{th} subject, on the p variables.

Let $x_i \equiv (x_{ji})$, $j = 1, \ldots, p$; $\mu \equiv (\mu_j)$, $j = 1, \ldots, p$; $\epsilon_i \equiv (\epsilon_{ji})$, $j = 1, \ldots, p$; and $f_i \equiv (f_{ki})$, $k = 1, \ldots, m$.

The model partitions or separates the observed score x_{ji} into a population mean μ_j, a "common" (to several questions) part f_i, and "specific" (to the questions) part ϵ_{ji}.

$$(x_{ji} | \mu_i, \lambda_j, f_i, m) = \mu_j + \lambda_j' \ f_i + \epsilon_{ji} \tag{2.1.2}$$

assuming that $\text{var}(\epsilon_i)$ is diagonal.

The factor analysis model is used when we wish to identify the number and the nature of the underlying factors responsible for covariation in the observables. This is also when the score of an individual is more related to his own scores than to the scores of other individuals. We wish to determine if there is an underlying set of m unobservable variables that describe the relationship between the p observable variables, where $m < p$.

We now consider the interpretation of the factor loading matrix. The covariance between the observations and the factor scores is

$$\begin{aligned}
\text{cov}(x_i, f_i) &= E(x_i f_i') - E(x_i)E(f_i') \\
&= E(\mu + \Lambda f_i) f_i' \\
&= E(\Lambda f_i f_i') \\
&= \Lambda E(f_i f_i')
\end{aligned}$$

$$= \Lambda R \tag{2.1.3}$$

Under the orthogonal factor model, $R = I_m$, and the factor loading matrix Λ is interpreted as a matrix of covariances (correlations) between the p observed scores and the m unobserved factors. The element in the j^{th} row and the k^{th} column is the covariance (correlation) between the j^{th} question and the k^{th} factor score. Thus a large element of Λ imply a strong relationship between the the corresponding question and factor score.

There are certain model assumptions that will be presented later that detail the rest of the model.

We can also write the model for all the observations in terms of matrices just as in regression. If we join the observation vectors into a matrix, then we can write the model as

$$\underset{(N \times p)}{(X|\mu, \Lambda, F, m)} = \underset{(N \times p)}{e \otimes \mu'} + \underset{(N \times m)}{F} \underset{(m \times p)}{\Lambda'} + \underset{(N \times p)}{E} , \tag{2.1.4}$$

where

$X =$ an $N \times p$ matrix of observed responses, $X' \equiv (x_1, \ldots, x_N)$,

$e =$ an N dimensional unit vector,

$\mu =$ a p-dimensional population mean vector,

$F =$ an $N \times m$ matrix of unobserved "common" factor scores, $F' = (f_1, \ldots, f_N)$,

$\Lambda =$ a $p \times m$ matrix of unknown constants called the factor loadings,

$E =$ an $N \times p$ matrix of "specific" errors or disturbance terms, $E = (\epsilon_1, \ldots, \epsilon_N)$,

and

$\otimes$ denotes the direct or Kroneker product.

If we stack the observation vectors into a single vector which is $Np \times 1$ then, we can write the model as

$$\underset{(Np \times 1)}{(x|\mu, \Lambda, f, m)} = \underset{(Np \times 1)}{e \otimes \mu} + \underset{(Np \times Nm)}{(I_N \otimes \Lambda)} \underset{(Nm \times 1)}{f} + \underset{(Np \times 1)}{\epsilon}, \tag{2.1.5}$$

where

$x =$ an Np-dimensional vector of observed responses, $x \equiv (x_1', \ldots, x_N')'$,

$e =$ an N dimensional unit vector,

$f =$ an Nm-dimensional unobserved "common" factor score vector, $f = (f_1', \ldots, f_N')'$, and

$\epsilon =$ an Np-dimensional "specific" error vector, $\epsilon = (\epsilon_1', \ldots, \epsilon_N')'$.

We will use all three representations of the model in equations 2.1.1, 2.1.4, and 2.1.5.

There have been many approaches to estimating the values of μ, Λ, and the f_i's. Early statistical factor analysis included maximum likelihood factor analysis (Lawley, 1940). In Lawley's model, the errors and the scores are assumed to be independent and normally distributed. The joint distribution of the scores and the observations is integrated with respect to the scores. Then maximum likelihood estimates are obtained for the loadings and the error covariance matrix. More recently the EM algorithm for maximum likelihood factor analysis (Rubin and Thayer, 1982) has been used. Again, independent normal distributions are assumed for the errors and factor scores. The scores are estimated in an E-step then the loadings and disturbance error variances are estimated in an M-step. These two methods lead us into the most recent methods.

Recently, Bayesian factor analysis methods have emerged (see Press & Shigemasu, 1989/1997; Lee, 1994; Lee, & Press, 1998; Hayashi, 1997; and Rowe & Press, 1998;) in which proper natural conjugate prior distributions are assessed for the unknown parameters Λ, the f_i's, and the covariance matrix Ψ. Bayesian posterior estimates are then obtained. These Bayesian methods are conditional on the number of factors. There has also been work (Press and Shigemasu, 1994) in Bayesian factor analysis in which the number of factors is an additional unknown parameter, a prior distribution is assessed for it, and a Bayesian marginal posterior estimate is obtained.

We will review the methods of maximum likelihood, EM, Press & Shigemasu 1989 (PS89), and Rowe & Press 1998 before presenting the new work developed in this dissertation. We will also include in an appendix a simple extension of PS89 and RP98 that places a prior distribution on a general observation mean.

2.2 Maximum Likelihood Factor Analysis

This is a review of the maximum likelihood factor analysis model (Lawley, 1940). We assume that the “specific” errors of the observations and the “common” factor scores are independent and normally distributed, then marginalize with respect to the scores, and maximize the resulting likelihood with respect to the loadings and the error or disturbance covariance matrix. We will make the following model assumptions.

Assumptions of the Model

For $i = 1, \ldots, N$:
(a) $\epsilon_i \sim N(0, \Psi)$, where $\Psi \equiv diag(\psi_1, \ldots, \psi_p)$ and $\psi_j > 0$, $j = 1, \ldots, p$;
(b) $(f_i|m) \sim N(0, R)$, $m \leq p$, and R usually taken to be I_m;
(c) ϵ_i and f_i are independent.

From (a)–(c) above, we see that the observations given the mean, the factor loadings, and the factor scores is normally distributed as expressed by

$$(x_i|\mu, \Lambda, f_i, m) \sim N(\mu + \Lambda f_i, \Psi),$$

the factor scores given their correlation matrix is normally distributed

$$(f_i|R, m) \sim N(0, R),$$

and the joint distribution of the factor scores with the observations is

$$p(f_i, x_i|\mu, \Lambda, R, \Psi, m) \propto e^{-\frac{1}{2}(f_i - \hat{f}_i)'(R^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}(f_i - \hat{f}_i)} e^{-\frac{1}{2}(x_i - \mu)'(\Psi + \Lambda R \Lambda')^{-1}(x_i - \mu)}$$

where

$$\hat{f}_i = (R^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}(x_i - \mu).$$

In this method, we find the marginal density of the data to be

$$\begin{aligned} p(x_i|\mu, \Lambda, R, \Psi, m) &= \int p(f_i, x_i|\mu, \Lambda, R, \Psi)\; df_i \\ &= (2\pi)^{-\frac{p}{2}}|\Lambda R\Lambda' + \Psi|^{-\frac{1}{2}}e^{-\frac{1}{2}(x_i-\mu)'(\Lambda R\Lambda'+\Psi)^{-1}(x_i-\mu)} \end{aligned}$$

For convenience, we will define the covariance matrix of this distribution to be Σ thus, $\Sigma = (\Lambda R\Lambda' + \Psi)$ and the distribution of $(x_i|\mu, \Lambda, R, \Psi, m)$ is

$$p(x_i|\mu, \Lambda, R, \Psi, m) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)}.$$

That is, $(x_i|\mu, \Lambda, R, \Psi, m)$ is normally distributed with mean μ and variance-covariance matrix $\Sigma = (\Lambda R\Lambda' + \Psi)$. For simplicity, we will assume the orthogonal factor model in which $R = I_m$.

The likelihood function L of the observations is

$$\begin{aligned} L &= p(x_1, \ldots, x_N|\mu, \Lambda, \Psi, m) \\ &= (2\pi)^{-\frac{Np}{2}}|\Sigma|^{-\frac{N}{2}}e^{-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)} \\ &= (2\pi)^{-\frac{Np}{2}}|\Sigma|^{-\frac{N}{2}}e^{-\frac{1}{2}trS\Sigma^{-1}} \qquad (2.2.1) \end{aligned}$$

where

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)'.$$

Maximizing the above likelihood with respect to μ yields $\hat{\mu} = \bar{x}$. We center the variables about $\bar{x}$ so that the variables will have a mean of zero. This eliminates the need for a mean μ in the model.

As we stated earlier, in order to avoid indeterminacies, we must add the side condition that $\Gamma = \Lambda'\Psi^{-1}\Lambda$ be a diagonal matrix. That is, without this condition, there will be an infinite number of solutions each related to the other by an orthogonal rotation (see Press 1982, p.339). Neglecting the terms that do not include Σ, the log likelihood is

$$LL = -\frac{N}{2}\left(log|\Sigma| + trS\Sigma^{-1}\right).$$

but maximizing LL with respect to Λ and Ψ is equivalent to minimizing

$$\begin{aligned} LL^* &= -\frac{2}{N}LL^* - log|S| - p \\ &= trS\Sigma^{-1} + log|S\Sigma^{-1}| - p. \end{aligned}$$

Upon differentiating LL^* with respect to Λ and Ψ, setting the result equal to a zero matrix, and applying some algebra we arrive at

$$(S - \hat{\Sigma})\hat{\Sigma}^{-1}\hat{\Lambda} = 0 \tag{2.2.2}$$

and

$$\hat{\Psi} = diag(S - \hat{\Lambda}\hat{\Lambda}'). \tag{2.2.3}$$

The above equations yield unique maximum likelihood estimators with the imposed side condition but there is not a closed form analytic solution; thus, they must be solved numerically.

To estimate the factor loadings, we can calculate the conditional distribution

$$p(f_i|x_i, \Lambda, \Psi, m) \propto e^{-\frac{1}{2}(f_i-\hat{f}_i)'(I_m+\Lambda'\Psi^{-1}\Lambda)(f_i-\hat{f}_i)}$$

where

$$\hat{f}_i = (I_m + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}x_i$$

and estimate the scores by $\hat{f}_i$ conditional on the observations, the loadings, and the error covariance matrix.

2.3 EM Maximum Likelihood Factor Analysis

The assumptions of EM maximum likelihood (Rubin and Thayer, 1982) are the same as those for maximum likelihood estimation in the previous section. The EM algorithm is simply a convenient method of obtaining maximum likelihood estimates. The algorithm consists of an E-step finding the expected value of the log likelihood for the scores F given the observed data X, then an M-step maximizing the expected log likelihood found in the E-step. This requires finding the expected value of the sufficient statistics

$$C_{xx} = \sum_{i=1}^{N} \frac{x_i x_i'}{N},$$

$$C_{xf} = \sum_{i=1}^{N} \frac{x_i f_i'}{N},$$

$$C_{ff} = \sum_{i=1}^{N} \frac{f_i f_i'}{N}$$

over F given X. The conditional mean of the scores is

$$E(f_i | x_i, \Psi, \Lambda, R, m) = \delta x_i$$

and the conditional covariance matrix of the scores is

$$Var(f_i | x_i, \Psi, \Lambda, R, m) = \Delta.$$

where

$$\begin{aligned}\delta &= (R^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}, \\ \Delta &= (R^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}. \end{aligned} \tag{2.3.1}$$

The conditional expectations of the sufficient statistics are

$$\begin{aligned} E(C_{xx}|X, \Psi, \Lambda, R, m) &= C_{xx}, \\ E(C_{xf}|X, \Psi, \Lambda, R, m) &= C_{xx}\delta, \\ E(C_{ff}|X, \Psi, \Lambda, R, m) &= \delta C_{xx}\delta' + \Delta. \end{aligned}$$

Upon maximizing the expected log likelihood found in the E-step we get

$$\begin{aligned} \Lambda &= (F'F)^{-1}F'X \\ \Psi &= diag\left\{\frac{(X - F\Lambda')'(X - F\Lambda')}{N}\right\}, \\ R &= \frac{F'F}{N}, \end{aligned} \tag{2.3.2}$$

if R is not taken to be the identity matrix.

We then start with initial values and cycle between (2.3.1) and (2.3.2) to obtain maximum likelihood estimates.

2.4 Bayesian Factor Analysis

The development of Bayesian factor analysis has been recent and brief. The Press and Shigemasu model is so far the best and most recent (see Lee, 1994 & Hyashi, 1997). Here is a description of the Bayesian factor analysis model of Press and Shigemasu, 1989 (PS89). In section 2.5 we will present the conditional modal estimation procedure of PS89, in section 2.6 we will present the LSO and Gibbs estimation procedures of RP98, and in an appendix we will present an extended Bayesian factor analysis model (EPS89 and ERP98) that places a prior distribution on the general mean μ.

We will write the factor analysis model in the vector representation of equation (2.1.5).

In PS89 an orthogonal model is assumed so $R = I_m$. Again, the model is

$$\underset{(Np \times 1)}{(x|\mu, \Lambda, f, m)} = \underset{(Np \times 1)}{e \otimes \mu} + \underset{(Np \times Nm)}{(I_N \otimes \Lambda)} \underset{(Nm \times 1)}{f} + \underset{(Np \times 1)}{\epsilon} . \tag{2.4.1}$$

where the variables are as defined before.

The estimate of the mean μ will be the mean of the observations $\bar{x}$ so to simplify things, PS89 assume that the observations have been centered about their sample mean resulting in a new mean of zero.

After the observations are centered about the sample mean the model becomes

$$\underset{(Np \times 1)}{(x|\Lambda, f, m)} = \underset{(Np \times Nm)}{(I_N \otimes \Lambda)} \quad \underset{(Nm \times 1)}{f} \quad + \quad \underset{(Np \times 1)}{\epsilon} \quad . \tag{2.4.2}$$

Likelihood

To obtain the likelihood for the PS89 model it is assumed that

(1) $\epsilon_i \sim N(0, \Psi)$,

where $\Psi > 0$ and $\mathrm{E}(\Psi)$ is diagonal to represent traditional beliefs of the model containing "common" and "specific" factors. This is analogous to assumption (a) of the maximum likelihood model but differs because here the assumption is that Ψ is a full positive definite matrix diagonal on average, not strictly a diagonal matrix. Writing $\epsilon = (\epsilon_1', \ldots, \epsilon_N')'$ we find that

$$\epsilon \sim N(0, I_N \otimes \Psi)$$

and thus the distribution of the observations is

$$[(x|f, \Lambda, \Psi, m) - (I_N \otimes \Lambda)f] \sim N(0, I_N \otimes \Psi)$$

which gives the likelihood for the observations as

$$p(x|f, \Lambda, \Psi, m) \propto |I_N \otimes \Psi|^{-\frac{1}{2}} e^{-\frac{1}{2}[x-(I_N \otimes \Lambda)f]'(I_N \otimes \Psi)^{-1}[x-(I_N \otimes \Lambda)f]} \tag{2.4.3}$$

where $\Psi > 0$ but diagonal on average. This likelihood for the observations can be written as the following matrix normal distribution

$$p(X|F,\Lambda,\Psi,m) \propto |\Psi|^{-\frac{N}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'(X-F\Lambda')} \quad \Psi > 0, \tag{2.4.4}$$

where the observations are $X' \equiv (x_1, \ldots, x_N)$ and the factor scores are $F' \equiv (f_1, \ldots, f_N)$.

We will use $p(\cdot)$ generically to denote "density"; the p's will be distinguished by their arguments.

Prior Distributions

In PS89, generalized natural conjugate families of prior distributions for the parameters are used. The factor loadings are assumed to depend on the disturbance covariance matrix. The disturbance covariance matrix is assumed to be independent of the factor scores. The factor scores are assumed to be independent of the factor loadings and the disturbance covariance matrix. The joint prior distribution for the parameters F, Λ, and Ψ is of the form:

$$p(F,\Lambda,\Psi|m) = p(\Lambda|\Psi,m)p(\Psi)p(F|m), \tag{2.4.5}$$

where

$$\begin{aligned} p(\Lambda|\Psi,m) &\propto |\Psi|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'}, & (2.4.6)\\ p(\Psi) &\propto |\Psi|^{-\frac{\nu}{2}} e^{-\frac{1}{2}tr\Psi^{-1}B}, \quad \nu > 2p, & (2.4.7)\\ p(F|m) &\propto e^{-\frac{1}{2}trF'F} & (2.4.8) \end{aligned}$$

with $\Psi > 0$, $H > 0$, and $B > 0$ a diagonal matrix (consequently $\mathrm{E}(\Psi)$ is diagonal). Thus, Λ conditional on Ψ has elements which are jointly normally distributed, and (Λ_0, H) are hyperparameters to be assessed; Ψ^{-1} follows a Wishart distribution, (ν, B) are hyperparameters to be assessed. The factor scores are independent and normally distributed. Note that $E(\Psi|B)$ is diagonal, to represent traditional views of the factor model containing "common" and "specific" factors.

Note that if $\Lambda' \equiv (\lambda_1, \ldots, \lambda_p)$, $\lambda \equiv vec(\Lambda') = (\lambda_1', \ldots, \lambda_p')'$, then $\mathrm{var}(\lambda|\Psi, m) = \Psi \otimes H^{-1}$, $\mathrm{var}(\lambda|m) = (E\Psi) \otimes H^{-1}$, and $\mathrm{cov}[(\lambda_i, \lambda_j)|\Psi, m] = \psi_{ij} H^{-1}$. Moreover, it is assumed that $H = n_0 I$, for some preassigned scalar n_0. These interpretations of the hyperparameters simplify assessment.

Also note that they have assumed

(2) $(f_i|m) \sim N(0, I_m)$.

This is the same as assumption (b) in the maximum likelihood model where $R = I_m$. Writing the factor scores as $f = (f_1', \ldots, f_N')'$ we find that

$$(f|m) \sim N(0, I_N \otimes I_m)$$

and the density of the factor scores can be written as the matrix normal distribution $p(F|m)$.

Also note that PS89 have assumed that

(3) ϵ_i and f_i are independent.

This is the same as assumption (c) of the maximum likelihood model. This assumption is evident from the likelihood and the prior distribution for the factor scores.

Posterior

By Bayes' rule, the posterior distribution for the unknown parameters of interest is

$$p(F, \Lambda, \Psi | X, m) \quad \propto \quad e^{-\frac{1}{2} trF'F} |\Psi|^{-\frac{(N+m+\nu)}{2}} e^{-\frac{1}{2} tr\Psi^{-1}U} \tag{2.4.9}$$

where

$$U \equiv (X - F\Lambda')'(X - F\Lambda') + (\Lambda - \Lambda_0)H(\Lambda - \Lambda_0)' + B.$$

2.5 Conditional Modal Estimation, PS89

In PS89, conditional modal estimation is used. The marginal mode $E(F|X) = \hat{F}$ is computed, then the conditional mode $E(\Lambda|\hat{F}, X) = \hat{\Lambda}$, and the modal value $\hat{\Psi}_{mode}$ given $\hat{\Lambda}$, $\hat{F}$, and X is found. Equation (2.4.9) is integrated with respect to the disturbance covariance matrix Ψ and the factor loadings Λ to obtain

$$p(F|X, m) \propto \frac{e^{-\frac{1}{2}trF'F}|H + F'F|^{\frac{\gamma-m-p}{2}}}{|A + (F - \hat{F})'(I_N - XW^{-1}X')(F - \hat{F})|^{\frac{N+m+\nu-p-1}{2}}}, \tag{2.5.1}$$

where we define the following

$$\begin{aligned}
\hat{F} &\equiv (I_N - XW^{-1}X')^{-1}XW^{-1}\Lambda_0 H \\
&= (I_N - X(X'X - W)^{-1}X')XW^{-1}\Lambda_0 H && (2.5.2)\\
W &\equiv X'X + B + \Lambda_0 H \Lambda_0', && (2.5.3)\\
A &\equiv H - (\Lambda_0 H)'W^{-1}\Lambda_0 H \\
&\quad -(XW^{-1}\Lambda_0 H)'(I_N - XW^{-1}X')^{-1}(XW^{-1}\Lambda_0 H) \\
&\equiv H - (\Lambda_0 H)'W^{-1}\Lambda_0 H \\
&\quad -(XW^{-1}\Lambda_0 H)'(I_N - X(X'X - W)^{-1}X')(XW^{-1}\Lambda_0 H). && (2.5.4)
\end{aligned}$$

When the sample size N is large, $F'F \approx NI_m$, by the weak law of large numbers. The two terms in the numerator can now be incorporated into the proportionality constant and the marginal posterior density of F becomes

$$p(F|X,m) \propto \frac{1}{|A + (F - \hat{F})'(I_N - XW^{-1}X')(F - \hat{F})|^{\frac{\gamma - m}{2}}}. \tag{2.5.5}$$

This is the kernel of a matrix T-distribution. The large sample posterior mean (and modal) estimator of F is $E(F|X,m) = \hat{F}$. We now estimate of Λ for given $F = \hat{F}$.

The conditional distribution of Λ for given F is

$$p(\Lambda|F,X,m) \propto \frac{1}{|R_F + (\Lambda - \Lambda_F)Q_F(\Lambda - \Lambda_F)'|^{-\frac{\gamma}{2}}}, \tag{2.5.6}$$

where

$$\begin{aligned} Q_F &= H + F'F, \\ R_F &= X'X + B + \Lambda_0 H \Lambda_0' - (X'F + \Lambda_0 H)Q_F^{-1}(X'F + \Lambda_0 H)', && (2.5.7) \\ \Lambda_F &= (X'F + \Lambda_0 H)(H + F'F)^{-1} && (2.5.8) \\ \gamma &= N + m + \nu - p - 1 && (2.5.9) \end{aligned}$$

That is $(\Lambda|F,X)$ follows a matrix T-distribution. Their posterior conditional mean (and modal) estimator of Λ is $E(\Lambda|\hat{F},X)$, or

$$\hat{\Lambda} = \Lambda_{\hat{F}} = (X'\hat{F} + \Lambda_0 H)(H + \hat{F}'\hat{F})^{-1}. \tag{2.5.10}$$

The covariance matrix Ψ is estimated conditional upon $(\Lambda, F) = (\hat{\Lambda}, \hat{F})$.

The conditional density of $(\Psi|\hat{\Lambda}, \hat{F}, X)$ is

$$p(\Psi|\hat{\Lambda}, \hat{F}, X, m) \propto \frac{e^{-\frac{1}{2} tr \Psi^{-1} \hat{U}}}{|\Psi|^{\frac{(N+m+\nu)}{2}}}, \quad \Psi > 0 \tag{2.5.11}$$

where

$$\hat{U} \;=\; (X - \hat{F}\hat{\Lambda}')'(X - \hat{F}\hat{\Lambda}') + (\hat{\Lambda} - \Lambda_0)H(\hat{\Lambda} - \Lambda_0)' + B. \tag{2.5.12}$$

The conditional posterior mean $E(\Psi|\hat{\Lambda}, \hat{F}, X, m)$ of $p(\Psi|\hat{\Lambda}, \hat{F}, X, m)$ is

$$\hat{\Psi} \;=\; \frac{\hat{U}}{N + m + \nu - 2p - 2}. \tag{2.5.13}$$

It should be noted that the conditional mode of $p(\Psi|\hat{\Lambda}, \hat{F}, X, m)$ is not the same as the conditional mean. The conditional mode is

$$\hat{\Psi}_{mode} \;=\; \frac{\hat{U}}{N + m + \nu}. \tag{2.5.14}$$

The estimators $(\hat{F}, \hat{\Lambda}, \hat{\Psi}_{mode})$ are conditional posterior modal estimators.

2.6 Computer Intensive Methods of Estimation, RP98

As previously stated, in PS89 a large sample approximation is used to estimate the factor scores. This large sample approximation can be avoided by using either Gibbs sampling or what we call Lindley/Smith optimization, (Rowe and Press, 1998 henceforth RP98). For a brief description of Lindley/Smith optimization henceforth LSO and Gibbs Sampling see appendix A.

Both Gibbs sampling and LSO require the posterior conditionals. Gibbs sampling requires the conditionals for the generation of random samples while LSO requires them for maximization by cycling through their modes. Rowe and Press find that Gibbs sampling is a better estimation procedure than LSO because Gibbs sampling through conditioning can yield both marginal point and interval estimates while LSO can yield only conditional point and interval estimates.

Conditional Posterior Densities

We find that the conditional posterior density of the factor scores is

$$\begin{aligned} p(F|\Lambda,\Psi,X,m) &\propto p(F,\Lambda,\Psi)p(X|F,\Lambda,\Psi) \\ &\propto p(F|m)p(\Lambda|\Psi)P(\Psi)p(X|F,\Lambda,\Psi) \\ &\propto e^{-\frac{1}{2}trF'F}|\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'(X-F\Lambda')} \\ &\propto e^{-\frac{1}{2}trF'F}e^{-\frac{1}{2}tr(X-F\Lambda')\Psi^{-1}(X-F\Lambda')'} \end{aligned}$$

which after some algebra can be written as

$$p(F|\Lambda, \Psi, X, m) \quad \propto \quad e^{-\frac{1}{2}tr(F-\tilde{F})(I_m+\Lambda'\Psi^{-1}\Lambda)(F-\tilde{F})'} \tag{2.6.1}$$

where $\tilde{F} \equiv X\Psi^{-1}\Lambda(I_m + \Lambda'\Psi^{-1}\Lambda)^{-1}$.

That is, the factor scores given the factor loadings, the disturbance covariance matrix, and the data is normally distributed.

The conditional posterior density of the factor loadings is

$$\begin{aligned}
p(\Lambda|F, \Psi, X, m) \quad &\propto \quad p(F, \Lambda, \Psi)p(X|F, \Lambda, \Psi) \\
&\propto \quad p(\Lambda|\Psi, m)p(\Psi)p(F|m)p(X|F, \Lambda, \Psi, m) \\
&\propto \quad |\Psi|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'} \\
&\cdot \quad |\Psi|^{-\frac{N}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'(X-F\Lambda')} \\
&\propto \quad e^{-\frac{1}{2}tr\Psi^{-1}[(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'+(X-F\Lambda')'(X-F\Lambda')]}
\end{aligned}$$

which after some algebra becomes

$$p(\Lambda|F, \Psi, X, m) \quad \propto \quad e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\tilde{\Lambda})(H+F'F)(\Lambda-\tilde{\Lambda})'} \tag{2.6.2}$$

where

$$\tilde{\Lambda} = (X'F + \Lambda_0 H)(H + F'F)^{-1}.$$

The conditional posterior density of the factor loadings given the factor scores, the disturbance covariance matrix, and the data is normally distributed.

The conditional posterior density of the disturbance covariance matrix is

$$\begin{aligned} p(\Psi|F,\Lambda,X,m) &\propto p(\Psi)p(\Lambda|\Psi,m)p(F|m)p(X|F,\Lambda,\Psi,m) \\ &\propto |\Psi|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Psi^{-1}B}|\Psi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'} \\ &\cdot |\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'(X-F\Lambda')} \\ &\propto |\Psi|^{-\frac{(N+m+\nu)}{2}}e^{-\frac{1}{2}tr\Psi^{-1}U} \end{aligned} \tag{2.6.3}$$

where

$$U = (X-F\Lambda')(X-F\Lambda')' + (\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)' + B.$$

That is, the conditional density of the disturbance covariance matrix given the factor scores, the factor loadings, and the data has an inverted Wishart density.

The modes of these conditional distributions are $\tilde{F}$, $\tilde{\Lambda}$ (as defined above), and

$$\tilde{\Psi} = \frac{U}{N+m+\nu}, \tag{2.6.4}$$

respectively.

Gibbs Sampling Estimation

For Gibbs estimation of the posterior, we start with initial values for F and Ψ say $\bar{F}_{(0)}$ and $\bar{\Psi}_{(0)}$. Then cycle through

$$\begin{aligned}
\bar{\Lambda}_{(i+1)} &\equiv \text{ a random sample from } p(\Lambda|\bar{F}_{(i)}, \bar{\Psi}_{(i)}, X) \\
\bar{\Psi}_{(i+1)} &\equiv \text{ a random sample from } p(\Psi|\bar{F}_{(i)}, \bar{\Lambda}_{(i+1)}, X) \\
\bar{F}_{(i+1)} &\equiv \text{ a random sample from } p(F|\bar{\Lambda}_{(i+1)}, \bar{\Psi}_{(i+1)}, X)
\end{aligned}$$

and we have $(\bar{\Lambda}_{(1)}, \bar{\Psi}_{(1)}, \bar{F}_{(1)}), \ldots, (\bar{\Lambda}_{(s)}, \bar{\Psi}_{(s)}, \bar{F}_{(s)})$, $(\bar{\Lambda}_{(s+1)}, \bar{\Psi}_{(s+1)}, \bar{F}_{(s+1)})$, ..., $(\bar{\Lambda}_{(s+t)}$, $\bar{\Psi}_{(s+t)}, \bar{F}_{(s+t)})$. The first s random samples called the "burn in" are discarded and the remaining t samples are kept. The means of the remaining random samples

$$\bar{F} = \frac{1}{t}\sum_{k=1}^{t} \bar{F}_{(s+k)}$$

$$\bar{\Lambda} = \frac{1}{t}\sum_{k=1}^{t} \bar{\Lambda}_{(s+k)}$$

$$\bar{\Psi} = \frac{1}{t}\sum_{k=1}^{t} \bar{\Psi}_{(s+k)}$$

are the sampling based posterior marginal mean estimates of the parameters.

LSO Estimation

For the hill climbing LSO estimation of the parameters (see appendix A), we start with an initial value for $\tilde{F}$, say $\tilde{F}_{(0)}$ then cycle through

$$\begin{aligned}
\tilde{\Lambda}_{(i+1)} &\equiv (X'\tilde{F}_{(i)} + \Lambda_0 H)(H + \tilde{F}'_{(i)}\tilde{F}_{(i)})^{-1} \\
\tilde{\Psi}_{(i+1)} &\equiv \frac{(X - \tilde{F}_{(i)}\tilde{\Lambda}'_{(i+1)})'(X - \tilde{F}_{(i)}\tilde{\Lambda}'_{(i+1)}) + (\tilde{\Lambda}_{(i+1)} - \Lambda_0)H(\tilde{\Lambda}_{(i+1)} - \Lambda_0)' + B}{N + m + \nu} \\
\tilde{F}_{(i+1)} &\equiv X\tilde{\Psi}^{-1}_{(i+1)}\tilde{\Lambda}_{(i+1)}(I_m + \tilde{\Lambda}'_{(i+1)}\tilde{\Psi}^{-1}_{(i+1)}\tilde{\Lambda}_{(i+1)})^{-1}.
\end{aligned}$$

until convergence is reached and we have the joint modal estimator for the unknown parameters $(\tilde{F}, \tilde{\Lambda}, \tilde{\Psi})$.

3 Correlated Bayesian Factor Analysis Model

3.1 Introduction

The following is an extension of the PS89 model for Bayesian factor analysis. In maximum likelihood factor analysis, the covariance matrix for the errors of the observations is assumed to be diagonal while in the Bayesian factor analysis model of PS89, it is assumed to be positive definite but diagonal on average. In both models, the error vectors are assumed to be independent. We assume general covariance matrices for the errors of the observations, the factor scores, and the factor loadings. We still assume that the errors are diagonal on average to represent traditional beliefs of "common" and "specific" factors.

The PS89 Bayesian factor analysis model is based on independence of observation vectors. If the observation vectors are not independent, and if the correlation between the observations is not taken into account, then the covariance matrix that is factor analyzed is improperly estimated.

As stated earlier, we use the Bayesian approach to factor analysis because the classical factor analysis model is inherently indeterminate and Bayesian methods with the incorporation of available proper prior information eliminates the problem of indeterminacies. But once the observation vectors are permitted to be dependent, more parameters are introduced.

We use proper prior information and take advantage of simplifications of the covariance structures to reduce the number of distinct covariance terms. The motivation for this is also because of the large number of distinct parameters that must be estimated and the computational requirements for all of the distinct parameters are unrealistic.

In both maximum likelihood and PS89 although a mean is assumed for the observations, its estimator is the mean of the observations, so for simplification, the observations are assumed to be centered about their mean. We do the same, but in an appendix include the development for placing a prior distribution on the fixed but unknown observation mean and estimating it along with the other parameters.

In maximum likelihood factor analysis the factor scores can be assumed to be correlated and in PS89 the factor scores are assumed to independent; both models assume that the factor score vectors are independent while we assume that the both the factor scores and the factor score vectors are correlated.

Also, the number of factors is assumed to be known in both maximum likelihood and in PS89 (although Press and Shigemasu have a development in which the number of factors is unknown see Press and Shigemasu 1994). We assume the number of factors to be unknown but fixed and place a prior distribution on it.

3.2 Model

The factor analysis model is

$$\underset{(Np\times 1)}{(x|\mu,m,\Lambda,f)} = \underset{(Np\times 1)}{\mu} + \underset{(Np\times Nm)}{(I_N\otimes\Lambda)} \underset{(Nm\times 1)}{f} + \underset{(Np\times 1)}{\epsilon}, \tag{3.2.1}$$

where

$N =$ the number of subjects or observation vectors,

$p =$ the number of variables measured on each subject or the dimension of the observation vectors.

It is assumed that N is much larger than p.

$x =$ an Np-dimensional vector of observed responses from the N subjects on p variables, $x \equiv (x_1', \ldots, x_N')'$,

$\mu =$ an Np-dimensional mean vector of the observed responses from the N subjects on p variables, $\mu \equiv (\mu_1', \ldots, \mu_N')'$,

$m =$ the number of factors, $m \leq p$,

$\Lambda =$ a matrix of factor loadings, $\Lambda' = (\lambda_1, \ldots, \lambda_p)$,

$f =$ an Nm-dimensional vector of unobservable variables for the N subjects on m factors called the factor score vector, $f = (f_1', \ldots, f_N')'$, and

$\epsilon =$ an Np-dimensional vector of the errors or disturbance terms for the observation vector of the N subjects, on the p variables, $\epsilon = (\epsilon_1', \ldots, \epsilon_N')'$. The error vectors, ϵ_i's are correlated.

3.3 Likelihood

Regarding the errors of the observations, we will assume that they are as follows

(I) $(\epsilon|\Omega) \sim N(0, \Omega)$,

where $\Omega > 0$, and we assume that Ω is diagonal on average to represent traditional views of the factor model containing "common" and "specific" factors. This is analogous to assumption (a) in the maximum likelihood model and assumption (1) in the PS89 Bayesian model. Recall that the maximum likelihood model assumes that the disturbance covariance matrix Ψ is diagonal while the PS89 model assumes that Ψ is a full general positive definite matrix and diagonal on average.

Thus from assumption (I) the distribution for the observations is

$$(x|\mu, \Omega, m, \Lambda, f) \sim N\left(\mu + (I_N \otimes \Lambda)f, \Omega\right),$$

and the likelihood for the observation vector is

$$p(x|\mu, \Omega, m, f, \Lambda) = (2\pi)^{-\frac{Np}{2}} |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}[x-\mu-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-\mu-(I_N\otimes\Lambda)f]}, \quad \Omega > 0, \quad (3.3.1)$$

where the covariance matrix is $\Omega = (\Omega_{ij})$, Ω_{ij} a $p \times p$ variance/covariance matrix.

As stated earlier, in order to keep from obfuscating the thrust of correlated Bayesian factor analysis, we will assume that each of the observations have the same mean μ, estimate it by the maximum likelihood estimate $\bar{x}$, and center the observations about the sample mean. In an appendix, we include the development for the general mean being a fixed but unknown parameter, place a prior distribution on it, and estimate it a posteriori along with the other parameters.

The model and likelihood now becomes

$$\underset{(Np\times 1)}{(x|m,\Lambda,f)} = \underset{(Np\times Nm)}{(I_N\otimes\Lambda)} \underset{(Nm\times 1)}{f} + \underset{(Np\times 1)}{\epsilon}, \tag{3.3.2}$$

and

$$p(x|\Omega,m,f,\Lambda) = (2\pi)^{-\frac{Np}{2}}|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]}. \tag{3.3.3}$$

We wish to obtain posterior estimates of the unknown parameters Ω, m, f, and Λ. Prior distributions will be assessed for the unknown parameters, their posterior distribution will be formed, and posterior estimators will be computed.

3.4 Priors

We will use natural conjugate prior distributions to represent our uncertainty about the parameters. We will assume that the joint prior distribution for the unknown parameters is given by

$$p(\Omega,m,f,\lambda) = p(\Omega)p(m)p(f|m)p(\lambda|m), \tag{3.4.1}$$

where

$$
\begin{aligned}
p(\Omega) &= c(Np,\nu)|\Omega|^{-\frac{\nu}{2}} e^{-\frac{1}{2} tr\Omega^{-1}A},\ \Omega > 0,\ \nu > 2Np, && (3.4.2)\\
p(m) &= \text{a discrete distribution to be defined below} && (3.4.3)\\
p(f|m) &= (2\pi)^{-\frac{Nm}{2}} |\Theta|^{-\frac{1}{2}} e^{-\frac{1}{2} f'\Theta^{-1} f},\quad \Theta > 0 && (3.4.4)\\
p(\lambda|m) &= (2\pi)^{-\frac{pm}{2}} |\Delta|^{-\frac{1}{2}} e^{-\frac{1}{2}(\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)},\quad \Delta > 0, && (3.4.5)
\end{aligned}
$$

and $c(Np, \nu)$ is a constant depending only on Np and ν. Note that ν is more than twice the product of the number of observations and their dimension.

We assume that a priori, the error disturbance covariance matrix is inverted Wishart where the hyperparameter matrix A is diagonal so that $E(\Omega)$ is diagonal. The factor scores and the factor loadings are assumed to be independent and normally distributed given the number of factors. The prior distribution for the number of factors will be reserved for assessment by the researcher.

We will assume without loss of generality, that the variance for the factor scores is unity so that Θ is a correlation matrix (see Press 1982 p. 331). Note that $\Lambda' \equiv (\lambda_1, \ldots, \lambda_p)$, and we have written $\lambda \equiv vec(\Lambda') = (\lambda_1', \ldots, \lambda_p')'$. (We will denote vectors using lower case and matrices as upper case letters.) Also note that we have made the following assumption regarding the distribution of the factor scores. We assume

(II) $(f|m) \sim N(0, \Theta)$,

this is analogous to assumption (b) in the maximum likelihood model and assumption (2) in the PS89 Bayesian model. It differs from both in that we allow the factor score vectors to be correlated. We also assume

(III) $(f|m)$ and $(\epsilon|\Omega)$ are independent random vectors.

This is the identical to assumption (c) of the maximum likelihood model and assumption (3) in the PS89 Bayesian model. This assumption is evident from the likelihood and the prior distribution for the factor scores.

3.5 Posterior

By Bayes' rule, the joint posterior distribution for the unknown parameters of interest is given by

$$\begin{aligned}
p(\Omega, m, f, \lambda|x) &\propto p(\Omega, m, f, \lambda)p(x|\Omega, m, f, \Lambda) \\
&\propto p(\Omega)p(f|m)p(\lambda|m)p(m)p(x|\Omega, m, f, \Lambda) \\
&\propto |\Omega|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Omega^{-1}A}(2\pi)^{-\frac{Nm}{2}}|\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}f'\Theta^{-1}f} \\
&\cdot (2\pi)^{-\frac{pm}{2}}|\Delta|^{-\frac{1}{2}}e^{-\frac{1}{2}(\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)} \\
&\cdot p(m)|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]} \\
&\propto p(m)(2\pi)^{-\frac{(N+p)m}{2}}|\Omega|^{-\frac{(\nu+1)}{2}}e^{-\frac{1}{2}tr\Omega^{-1}A} \\
&\cdot |\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}f'\Theta^{-1}f}|\Delta|^{-\frac{1}{2}}e^{-\frac{1}{2}(\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)} \\
&\cdot e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]} \qquad (3.5.1)
\end{aligned}$$

In the next section we will derive the posterior conditional distributions.

3.6 Conditional Posterior Densities

Here we find the posterior conditional densities needed for Gibbs sampling.

We find that the conditional posterior density for the error covariance matrix is

$$\begin{aligned}
p(\Omega|m,f,\lambda,x) &\propto p(\Omega,m,f,\lambda)p(x|\Omega,m,f,\lambda) \\
&\propto p(\Omega)p(m)p(f|m)p(\lambda|m)p(x|\Omega,m,f,\lambda) \\
&\propto p(\Omega)p(x|\Omega,m,f,\lambda) \\
&\propto |\Omega|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Omega^{-1}A} \\
&\cdot \ |\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]} \\
&\propto |\Omega|^{-\frac{(\nu+1)}{2}}e^{-\frac{1}{2}tr\Omega^{-1}U} \qquad (3.6.1)
\end{aligned}$$

where

$$U = [x-(I_N\otimes\Lambda)f][x-(I_N\otimes\Lambda)f]' + A.$$

That is, the conditional posterior density of the error covariance matrix given the number of factors, the factor scores, the factor loadings, and the data is an inverted Wishart.

We find the conditional posterior density of the factor scores as follows

$$\begin{aligned}
p(f|\Omega,m,\lambda,x) &\propto p(\Omega,m,f,\lambda)p(x|\Omega,m,f,\lambda) \\
&\propto p(\Omega)p(m)p(f|m)p(\lambda|m)p(x|\Omega,m,f,\lambda) \\
&\propto p(f|m)p(x|\Omega,m,f,\lambda) \\
&\propto (2\pi)^{-\frac{Nm}{2}}|\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}f'\Theta^{-1}f}
\end{aligned}$$

$$
\begin{aligned}
\cdot &\quad |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]} \\
\propto &\quad e^{-\frac{1}{2}(f-\tilde{f})'\left[\Theta^{-1}+(I_N\otimes\Lambda)'\Omega^{-1}(I_N\otimes\Lambda)\right](f-\tilde{f})} \qquad (3.6.2)
\end{aligned}
$$

where

$$\tilde{f} = \left[\Theta^{-1} + (I_N \otimes \Lambda)'\Omega^{-1}(I_N \otimes \Lambda)\right]^{-1} (I_N \otimes \Lambda)'\Omega^{-1}x.$$

The factor scores given the error covariance matrix, the number of factors, the factor loadings, and the data follows a normal distribution.

We find the conditional posterior density of the factor loadings as follows

$$
\begin{aligned}
p(\lambda|\Omega, m, f, x) &\propto p(\Omega, m, f, \lambda)p(x|\Omega, m, f, \lambda) \\
&\propto p(\Omega)p(m)p(f|m)p(\lambda|m)p(x|\Omega, m, f, \lambda) \\
&\propto p(\lambda|m)p(x|\Omega, m, f, \lambda) \\
&\propto (2\pi)^{-\frac{pm}{2}}|\Delta|^{-\frac{1}{2}} e^{-\frac{1}{2}(\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)} \\
&\cdot\quad |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]} \\
&\propto e^{-\frac{1}{2}\gamma} \qquad (3.6.3)
\end{aligned}
$$

where

$$\gamma = (\lambda - \lambda_0)'\Delta^{-1}(\lambda - \lambda_0) + [x - (I_N \otimes \Lambda)f]'\Omega^{-1}[x - (I_N \otimes \Lambda)f]$$

and

$$\lambda = vec(\Lambda').$$

The factor loadings given the error covariance matrix, the number of factors, the factor scores, and the data does not follow a recognizable distribution. Thus, random samples must be generated by a multivariate rejection sampling technique.

The conditional posterior density of the number of factors is

$$\begin{aligned}
p(m|\Omega, f, \lambda, x) &\propto p(\Omega, m, f, \lambda)p(x|\Omega, m, f, \lambda) \\
&\propto p(\Omega)p(m)p(f|m)p(\lambda|m)p(x|\Omega, m, f, \lambda) \\
&\propto p(m)p(f|m)p(\lambda|m)p(x|\Omega, m, f, \lambda) \\
&\propto p(m)(2\pi)^{-\frac{(N+p)m}{2}}|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]} \\
&\cdot \quad |\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}f'\Theta^{-1}f}|\Delta|^{-\frac{1}{2}}e^{-\frac{1}{2}(\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)} \\
&\propto p(m)(2\pi)^{-\frac{(N+p)m}{2}}|\Omega|^{-\frac{1}{2}}|\Theta|^{-\frac{1}{2}}|\Delta|^{-\frac{1}{2}}e^{-\frac{1}{2}\tau} \qquad (3.6.4)
\end{aligned}$$

where

$$\tau = [x - (I_N \otimes \Lambda)f]'\Omega^{-1}[x - (I_N \otimes \Lambda)f] + f'\Theta^{-1}f + (\lambda - \lambda_0)'\Delta^{-1}(\lambda - \lambda_0)$$

and

$$\lambda = vec(\Lambda').$$

This is not a recognizable distribution regardless of our choice of prior distributions for the number of factors. This conditional posterior distribution depends on the number of factors in a complicated fashion. The dimension of several of the matrices depends on the number of factors.

3.7 Gibbs Sampling Estimation

We will use Gibbs sampling for estimation of the parameters in (3.5.1) because we can obtain marginal posterior point and interval estimates. We cannot use conditional modal estimation or LSO. Conditional modal estimation requires the posterior distribution to be integrated with respect to one of the parameters, which cannot be done in a closed form. LSO requires the conditional posterior distributions to be unimodal (to converge to a single mode) which is not always the case and LSO does not yield marginal point and interval estimates.

For Gibbs estimation of the posterior, we start with initial values for Ω, m, f, and λ say $\bar{\Omega}_{(0)}$, $\bar{m}_{(0)}$, $\bar{f}_{(0)}$, and $\bar{\lambda}_{(0)}$.

Then for a given number of factors $m = \bar{m}_{(i)}$ cycle through

$$\begin{aligned}
\bar{\Omega}_{(i+1)} &\equiv \text{ a random sample from } p(\Omega|\bar{f}_{(i)}, \bar{\lambda}_{(i)}, \bar{m}_{(i)}, x) \\
\bar{f}_{(i+1)} &\equiv \text{ a random sample from } p(f|\bar{\Omega}_{(i+1)}, \bar{\lambda}_{(i)}, \bar{m}_{(i)}, x) \\
\bar{\lambda}_{(i+1)} &\equiv \text{ a random sample from } p(\lambda|\bar{\Omega}_{(i+1)}, \bar{f}_{(i+1)}, \bar{m}_{(i)}, x).
\end{aligned}$$

which is the Gibbs sampling algorithm.

For the given number of factors $m = \bar{m}_{(i)}$ we have the sequence

$$
\begin{gathered}
(\bar{\lambda}_{(1)}, \bar{\Omega}_{(1)}, \bar{f}_{(1)}) \\
\vdots \\
(\bar{\lambda}_{(s)}, \bar{\Omega}_{(s)}, \bar{f}_{(s)}) \\
(\bar{\lambda}_{(s+1)}, \bar{\Omega}_{(s+1)}, \bar{f}_{(s+1)}) \\
\vdots \\
(\bar{\lambda}_{(s+t)}, \bar{\Omega}_{(s+t)}, \bar{f}_{(s+t)}).
\end{gathered}
$$

The first s random samples called the “burn in” are discarded and the remaining t samples are kept. The means of the remaining random samples

$$\bar{\Omega} = \frac{1}{t}\sum_{k=1}^{t} \bar{\Omega}_{(s+k)} \tag{3.7.1}$$

$$\bar{f} = \frac{1}{t}\sum_{k=1}^{t} \bar{f}_{(s+k)} \tag{3.7.2}$$

$$\bar{\lambda} = \frac{1}{t}\sum_{k=1}^{t} \bar{\lambda}_{(s+k)} \tag{3.7.3}$$

are the sampling based marginal posterior mean estimates of the parameters given the number of factors $m = \bar{m}_{(i)}$. We do this for each value of m, then find the value of the number of factors $m = \bar{m}$ that makes the posterior conditional distribution for the number of factors $p(m|\bar{\Omega}, \bar{f}, \bar{\lambda}, x)$ a maximum given the corresponding estimates of the other parameters. This is the same as finding the value for the number of factors that gives the largest conditional posterior odds ratio. We will have $(\bar{m}, \bar{\Omega}, \bar{f}, \bar{\lambda})$ as our posterior estimates of the unknown parameters where $(\bar{\Omega}, \bar{f}, \bar{\lambda})$ are the estimates conditional on $m = \bar{m}$.

It should be noted that in the posterior conditional distribution for the factor loadings $p(\lambda|\Omega, m, f, x)$, the terms in the exponent do not combine nicely to form a well known and recognizable distribution. Because of this, in order to draw a random sample from the conditional posterior distribution for λ requires a multivariate rejection sampling technique (see Gilks and Wild, 1992). This is an extremely difficult task, is very computer intensive, and is time consuming.

4 Covariance Simplifications and Motivation

The above is a very general model which requires many parameters to be estimated, an extremely difficult multivariate rejection sampling technique, and enormous storage requirements.

4.1 Covariance Simplifications

We simply the model by assuming structures for the covariance matrices of the loadings, the observations, and the factor scores.

We will specify that

$$\begin{aligned} var(\lambda|\Psi) &= \Delta \\ &= \Psi \otimes H^{-1} \end{aligned}$$

where H is a covariance hyperparameter to be assessed. This specification, first used in PS89 simplifies the covariance structure for the factor loadings. We will also specify that the error covariance and factor score structures are either

$$\Omega = \Phi \otimes \Psi, \quad \Phi > 0, \ \Psi > 0 \tag{4.1.1a}$$

and

$$\Theta = \Phi \otimes R, \quad \Phi > 0, \ R > 0 \tag{4.1.1b}$$

both separable covariance matrices where $\otimes$ denotes the Kroneker product, or

$$\Omega = \begin{pmatrix} \Psi & \Upsilon & & \cdots & \Upsilon \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \Upsilon \\ & & & & \Psi \end{pmatrix}, \quad \Psi > 0, \; \Upsilon > 0 \tag{4.1.2a}$$

and

$$\Theta = \begin{pmatrix} R & P & & \cdots & P \\ & R & & & \\ & & \ddots & & \vdots \\ & & & & P \\ & & & & R \end{pmatrix}, \quad R > 0, \; P > 0 \tag{4.1.2b}$$

both matrices with intraclass covariance/correlation structure. Doing so not only reduces the number of parameters that we must estimate and the enormous storage requirement, but as we will see, the posterior conditional distribution for the factor loadings is a nice recognizable distribution thus eliminating the need for a multivariate rejection sampling technique.

4.2 Covariance Motivation

Here we will present the two observation error covariance simplifications, along with further simplifications. The number of distinct parameters for each is given in Table 1 and the storage requirements are given in Table 2. In addition to the reduction in the number of parameters, the computations are simpler. The Gibbs sampling algorithm before simplification requires inversion, Cholesky factorization, and determinants of very large matrices as well as a multivariate rejection sampling technique for each iteration.

General Covariance Matrix

The full error covariance matrix for the observation vector is given by

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} & & \cdots & \Omega_{1N} \\ & \Omega_{22} & & & \\ & & \ddots & & \vdots \\ & & & & \Omega_{N-1,N} \\ & & & & \Omega_{NN} \end{pmatrix}.$$

where the variance of observation vector i is given by the $p \times p$ matrix

$$var(x_i|\Omega, m, f, \lambda) = \Omega_{ii}$$

and the covariance between observation vectors i and j is given by the $p \times p$ matrix

$$cov(x_i, x_j|\Omega, m, f, \lambda) = \Omega_{ij}.$$

Separable Covariance Matrix

If we can specify the following separable structure as our covariance matrix for the observations

$$\Omega \equiv \begin{pmatrix} \phi_{11}\Psi & \phi_{12}\Psi & & \cdots & \phi_{1N}\Psi \\ & \phi_{22}\Psi & & & \\ & & \ddots & & \vdots \\ & & & & \\ & & & & \phi_{NN}\Psi \end{pmatrix},$$

that is, the variance of the observation vectors are related by a multiplicative constant and given by the form

$$var(x_i|\Phi, \Psi, m, f, \lambda) = \phi_{ii}\Psi$$

and the covariance between any two observation vectors are related by a multiplicative constant

$$cov(x_i, x_j|\Phi, \Psi, m, f, \lambda) = \phi_{ij}\Psi,$$

then as previously stated, the conditional posterior distribution for the factor loadings has a convenient mathematical form given our choice of prior distributions.

If we can further identify the existence of the same variance for all of the observation vectors (homoscedasticity), then the covariance matrix becomes

$$\Omega \equiv \begin{pmatrix} \Psi & \phi_{12}\Psi & & \cdots & \phi_{1N}\Psi \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \\ & & & & \Psi \end{pmatrix}.$$

If the $\phi'_{ij}s$ depend only on a small number of parameters say $\rho = (\rho_1, \ldots, \rho_K)$, then the covariance matrix is

$$\Omega \equiv \begin{pmatrix} \Psi & \phi_{12}(\rho)\Psi & & \cdots & \phi_{1N}(\rho)\Psi \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \\ & & & & \Psi \end{pmatrix}$$

and the number of distinct parameters is further reduced. We will write the matrix Ω as $\Phi \otimes \Psi$ where we call Φ the between observation correlation matrix and Ψ the within observation covariance matrix. For simplicity, we will assume that ρ is a scalar. This could easily be extended.

Matrix Intraclass Covariance Matrix

If we can specify the matrix intraclass covariance error structure

$$\Omega = \begin{pmatrix} \Psi & \Upsilon & & \cdots & \Upsilon \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \Upsilon \\ & & & & \Psi \end{pmatrix}, \quad \Psi > 0, \ \Upsilon > 0. \qquad (4.2.1)$$

That is, the variance of any observation vector i is

$$var(x_i | \Psi, \Upsilon, m, f, \lambda) = \Psi$$

and the covariance between two observation vectors i and j is

$$cov(x_i, x_j | \Psi, \Upsilon, m, f, \lambda) = \Upsilon,$$

then we not only simplify the covariance structure and reduce the number of parameters that must be estimated, but as we will see, we will not need a multivariate rejection sampling technique.

	Ω_{ij}	$\Omega_{ij} = \phi_{ij}\Psi$	$\Omega_{ii} = \Psi$ $\Omega_{ij} = \phi_{ij}\Psi$	$\Omega_{ii} = \Psi$ $\Omega_{ij} = \phi_{ij}(\rho)\Psi$	$\Omega_{ii} = \Psi$ $\Omega_{ij} = 0$	$\Omega_{ii} = \Psi$ $\Omega_{ij} = \Upsilon$
(N, p)	$\frac{Np(Np+1)}{2}$	$\frac{N(N+1)}{2}$ $+\frac{p(p+1)}{2}$	$\frac{N(N+1)}{2} - N$ $+\frac{p(p+1)}{2}$	$1 + \frac{p(p+1)}{2}$	$\frac{p(p+1)}{2}$	$2\frac{p(p+1)}{2}$
(100,10)	500,500	5,105	5,005	56	55	110
(48,15)	259,560	1,296	1,248	121	120	240
(50,12)	180,300	1,353	1,303	79	78	156
(55,10)	151,525	1,585	1,540	56	55	110

Table 1: Number of Distinct Parameters in the Covariance Matrix.

In addition to the enormous number of distinct parameters, we may be limited by computer storage requirements. It is well known that 32 bits is used to represent numbers in a computer in single precision, there are eight bits to a byte, and that there are 1024 bytes in a megabyte. The formula for the number of megabytes to store the distinct parameters in single precision is

$$\begin{aligned} Number\ of\ Megabytes &= (Number\ of\ distinct\ parameters) \\ &\cdot \left(\frac{32\ bits}{distinct\ parameter}\right)\left(\frac{1\ byte}{8\ bit}\right)\left(\frac{1\ megabyte}{1024\ bytes}\right) \end{aligned}$$

and the numbers are given in Table 2.

			$\Omega_{ii} = \Psi$	$\Omega_{ii} = \Psi$	$\Omega_{ii} = \Psi$	$\Omega_{ii} = \Psi$
(N,p)	Ω_{ij}	$\Omega_{ij} = \phi_{ij}\Psi$	$\Omega_{ij} = \phi_{ij}\Psi$	$\Omega_{ij} = \phi_{ij}(\rho)\Psi$	$\Omega_{ij} = 0$	$\Omega_{ij} = \Upsilon$
(100,10)	1955.09	19.94	19.55	0.22	0.21	0.43
(48,15)	1013.91	5.06	4.88	0.47	0.47	0.94
(50,12)	704.30	5.29	5.09	0.31	0.30	0.61

Table 2: Number of Megabytes to Store Distinct Covariance Parameters.

5 Separable Covariance Models

As stated earlier, if we can specify a separable covariance matrix for the observation vector $x = (x_1', \cdots, x_N')$, then $var(x|\Phi, \Psi, m, f, \Lambda) = \Phi \otimes \Psi$. Separable covariance structures are used very often, for example in time series analysis. With a separable covariance matrix, the covariance between the i^{th} and j^{th} rows of X is $\phi_{ij}\Psi$, and the covariance between the i^{th} and j^{th} columns of X is $\psi_{ij}\Phi$ where as before $X' = (x_1, \ldots, x_N)$.

It is claimed that this covariance structure is equivalent to the assumption of [weak or covariance] stationarity used for time series (Basilevsky 1994, p. 489). We will call the matrix Φ the between observation covariance matrix and Ψ is the within observation covariance matrix.

5.1 Separable Model

Likelihood

The likelihood function under separability ($\Omega = \Phi \otimes \Psi$) is given by

$$p(x|\Phi, \Psi, m, f, \Lambda) = (2\pi)^{-\frac{Np}{2}} |\Phi \otimes \Psi|^{-\frac{1}{2}} e^{-\frac{1}{2}[x-(I_N \otimes \Lambda)f]'(\Phi \otimes \Psi)^{-1}[x-(I_N \otimes \Lambda)f]} \quad (5.1.1)$$

where the covariance matrix is given by

$$\Omega \equiv \Phi \otimes \Psi = \begin{pmatrix} \phi_{11}\Psi & \phi_{12}\Psi & & \cdots & \phi_{1N}\Psi \\ & \phi_{22}\Psi & & & \\ & & \ddots & & \vdots \\ & & & & \\ & & & & \phi_{NN}\Psi \end{pmatrix} \tag{5.1.2}$$

and the matrices $\Phi, \Psi > 0$ where $\mathrm{E}(\Psi)$ is diagonal. This likelihood for the observations can be simplified and rewritten as

$$p(X|\Phi, \Psi, m, F, \Lambda) = (2\pi)^{-\frac{Np}{2}} |\Phi|^{-\frac{p}{2}} |\Psi|^{-\frac{N}{2}} e^{-\frac{1}{2} tr \Psi^{-1} (X - F\Lambda')' \Phi^{-1} (X - F\Lambda')}. \tag{5.1.3}$$

where $X' \equiv (x_1, \ldots, x_N)$, $F' \equiv (f_1, \ldots, f_N)$, and using the facts that for $A > 0$ an $n_1 \times n_1$ matrix and $B > 0$ an $n_2 \times n_2$ matrix

$$|A \otimes B|^{-\frac{1}{2}} = |A|^{-\frac{n_2}{2}} |B|^{-\frac{n_1}{2}}$$

and

$$vec'(u - v)'(A \otimes B)^{-1} vec(u - v) = tr B^{-1} (U - V)' A^{-1} (U - V).$$

where $u = vec(U')$ and $v = vec(V')$.

Prior distributions will be assessed for the unknown parameters $(\Psi, \Phi, m, F, \Lambda)$.

Priors

We will use the same generalized natural conjugate families of prior distributions for the parameters as in Section 3. The prior distributions are simplified from their previous form by the adoption of the separable covariances. The joint prior distribution is given by

$$p(\Phi, \Psi, m, F, \Lambda) = p(\Psi)p(\Phi)p(m)p(F|\Phi, m)p(\Lambda|\Psi, m) \tag{5.1.4}$$

where

$$
\begin{aligned}
p(\Psi) &= c(p,\nu)|\Psi|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Psi^{-1}B}, && (5.1.5)\\
p(F|\Phi, m) &= (2\pi)^{-\frac{Nm}{2}}|R|^{-\frac{N}{2}}|\Phi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Phi^{-1}FR^{-1}F'}, && (5.1.6)\\
p(\Lambda|\Psi, m) &= (2\pi)^{-\frac{pm}{2}}|H^{-1}|^{-\frac{p}{2}}|\Psi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'}, && (5.1.7)
\end{aligned}
$$

$c(p,\nu)$ is a constant that depends only on p and ν, and the hyperparameters ν, B, Λ_0, H, those for $p(\Phi)$ and $p(m)$ are assessed as in appendix C. We assume that B is diagonal and consequently $\mathrm{E}(\Psi)$ is diagonal to represent traditional views of the model containing "common" and "specific" factors.

The prior distributions for the unknown parameters of interest are normal for the factor scores, and the factor loadings while it is inverted Wishart for the error disturbance covariance matrix.

We will discuss prior distributions for Φ later and the prior distribution for m is a discrete distribution to be assessed by the researcher.

Posterior

By Bayes' Rule, the joint posterior distribution for the unknown parameters of interest is given by

$$\begin{aligned} p(\Phi, \Psi, m, F, \Lambda|X) &\propto p(m)p(\Phi)(2\pi)^{-\frac{(N+p)m}{2}}|\Phi|^{-\frac{(p+m)}{2}}|H|^{\frac{p}{2}}|\Psi|^{-\frac{(N+m+\nu)}{2}} \\ &\cdot |R|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Phi^{-1}FR^{-1}F'}e^{-\frac{1}{2}tr\Psi^{-1}U} \end{aligned} \tag{5.1.8}$$

where the posterior conditional mean is given by

$$U \equiv (X - F\Lambda')'\Phi^{-1}(X - F\Lambda') + (\Lambda - \Lambda_0)H(\Lambda - \Lambda_0)' + B. \tag{5.1.9}$$

In Section 5.2 we will consider the case where Φ is known (for example $\Phi = I_N$ for independent observations), derive the necessary posterior conditionals, and discuss parameter estimation by Gibbs sampling. In Section 5.3, we will consider the case where Φ is a completely unknown general covariance matrix, derive the necessary posterior conditionals, and discuss parameter estimation by Gibbs sampling. In Section 5.4, we will consider the case where Φ is unknown, but structured so that it depends on only one parameter ρ, derive the necessary posterior conditionals, and discuss parameter estimation by Gibbs sampling.

Conditional Posterior Densities

From the joint posterior distribution we can obtain the posterior conditional distributions. The conditional posterior distribution of the disturbance covariance matrix is

$$\begin{aligned}
p(\Psi|\Phi, m, F, \Lambda, X) &\propto p(\Phi, \Psi, m, F, \Lambda)p(X|\Phi, \Psi, m, F, \Lambda) \\
&\propto p(\Phi)p(\Psi)p(m)p(F|\Phi, m)p(\Lambda|\Psi, m) \\
&\cdot\ p(X|\Phi, \Psi, m, F, \Lambda) \\
&\propto p(\Psi)p(\Lambda|\Psi, m)p(X|\Phi, \Psi, m, F, \Lambda) \\
&\propto |\Psi|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Psi^{-1}B}|\Psi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'} \\
&\cdot\ |\Phi|^{-\frac{p}{2}}|\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'\Phi^{-1}(X-F\Lambda')} \\
&\propto |\Psi|^{-\frac{(N+m+\nu)}{2}}e^{-\frac{1}{2}tr\Psi^{-1}U} \qquad (5.1.10)
\end{aligned}$$

where

$$U = (X - F\Lambda')'\Phi^{-1}(X - F\Lambda') + (\Lambda - \Lambda_0)H(\Lambda - \Lambda_0)' + B. \qquad (5.1.11)$$

The distribution of the disturbance covariance matrix given the correlation matrix, the number of factors, the factor scores, the factor loadings, and the data is an inverted Wishart.

The conditional posterior distribution for the factor scores is

$$\begin{aligned}
p(F|\Phi,\Psi,m,\Lambda,X) &\propto p(\Phi,\Psi,m,F,\Lambda)p(X|\Phi,\Psi,m,F,\Lambda) \\
&\propto p(\Phi)p(\Psi)p(m)p(F|\Phi,m)p(\Lambda|\Psi,m) \\
&\cdot\ p(X|\Phi,\Psi,m,F,\Lambda) \\
&\propto p(F|\Phi,m)p(X|\Phi,\Psi,m,F,\Lambda) \\
&\propto |\Phi|^{-\frac{m}{2}}|R|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Phi^{-1}FR^{-1}F'} \\
&\cdot\ |\Phi|^{-\frac{p}{2}}|\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'\Phi^{-1}(X-F\Lambda')} \\
&\propto |\Phi|^{-\frac{(p+m)}{2}}|\Psi|^{-\frac{N}{2}}|R|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Phi^{-1}[FR^{-1}F'+(X-F\Lambda')\Psi^{-1}(X-F\Lambda')']}
\end{aligned}$$

which after some algebra can be written as

$$p(F|\Phi,\Psi,m,\Lambda,X) \propto e^{-\frac{1}{2}tr\Phi^{-1}(F-\tilde{F})(R^{-1}+\Lambda'\Psi^{-1}\Lambda)(F-\tilde{F})'}$$

where the posterior conditional mean is given by

$$\tilde{F} = X\Psi^{-1}\Lambda(R^{-1}+\Lambda'\Psi^{-1}\Lambda)^{-1}.$$

The conditional posterior distribution for the factor scores given the correlation matrix, the disturbance covariance matrix, the number of factors, the factor loadings, and the data is normally distributed.

The conditional posterior density for the factor scores is

$$\begin{aligned}
p(\Lambda|\Phi,\Psi,m,F,X) &\propto p(\Phi,\Psi,m,F,\Lambda)p(X|\Phi,\Psi,m,F,\Lambda)\\
&\propto p(\Phi)p(\Psi)p(m)p(F|\Phi,m)p(\Lambda|\Psi,m)\\
&\cdot\ p(X|\Phi,\Psi,m,F,\Lambda)\\
&\propto p(\Lambda|\Psi,m)p(X|\Phi,\Psi,m,F,\Lambda)\\
&\propto |\Psi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'}\\
&\cdot\ |\Phi|^{-\frac{p}{2}}|\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'\Phi^{-1}(X-F\Lambda')}\\
&\propto e^{-\frac{1}{2}tr\Psi^{-1}[(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'+(X-F\Lambda')'\Phi^{-1}(X-F\Lambda')]}
\end{aligned}$$

which after some algebra becomes

$$p(\Lambda|\Phi,\Psi,m,F,X) \propto e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\tilde{\Lambda})(H+F'\Phi^{-1}F)(\Lambda-\tilde{\Lambda})'} \qquad (5.1.12)$$

where the posterior conditional mean is given by

$$\tilde{\Lambda}=[X'\Phi^{-1}F+\Lambda_0H](H+F'\Phi^{-1}F)^{-1}.$$

The conditional distribution for the factor scores given the correlation matrix, the disturbance covariance matrix, the number of factors, the factor scores, and the data is normally distributed.

All of these conditional posterior densities are well known recognizable distributions that do not require rejection sampling. Standard random variable generation methods can be used.

However, the conditional distribution for the number of factors is not tractable and recognizable.

The conditional posterior distribution of the number of factors is

$$\begin{aligned}
p(m|\Phi,\Psi,F,\Lambda,X) &\propto p(X|\Phi,\Psi,m,F,\Lambda)p(\Phi,\Psi,m,F,\Lambda) \\
&\propto p(X|\Phi,\Psi,m,F,\Lambda)p(\Phi)p(\Psi)p(F|\Phi,m)p(\Lambda|\Psi,m)p(m) \\
&\propto (2\pi)^{-\frac{Np}{2}}|\Phi|^{-\frac{p}{2}}|\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'\Phi^{-1}(X-F\Lambda')} \\
&\cdot\ p(\Phi)|\Psi|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Psi^{-1}B} \\
&\cdot\ (2\pi)^{-\frac{Nm}{2}}|R|^{-\frac{N}{2}}|\Phi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Phi^{-1}FR^{-1}F'} \\
&\cdot\ (2\pi)^{-\frac{pm}{2}}|H^{-1}|^{-\frac{p}{2}}|\Psi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'}p(m) \\
&\propto p(m)p(\Phi)(2\pi)^{-\frac{(N+p)m}{2}}|R|^{-\frac{N}{2}}|H|^{\frac{p}{2}} \\
&\cdot\ |\Phi|^{-\frac{(p+m)}{2}}|\Psi|^{-\frac{(N+m+\nu)}{2}}e^{-\frac{1}{2}\tau} \qquad (5.1.13)
\end{aligned}$$

where

$$\tau = trR^{-1}F'\Phi^{-1}F + tr\Psi^{-1}U.$$

As previously stated, the conditional posterior density for the number of factors given the correlation matrix, the disturbance covariance matrix, the factor scores, the factor loadings,and the data does not have a tractable and recognizable form.

We do not need to generate random samples from the conditional posterior distribution because we compute Gibbs sampling estimates for the parameters of interest for a given number of factors. We do this for each of the possible numbers of factors and compute the conditional posterior density of the number of factors given these Gibbs sampling estimates. We select the number of factors to be that value that maximizes the posterior conditional density. This is the same as selecting the number of factors to be the value with the largest conditional posterior odds ratio.

5.2 Separable Model, Φ Known

In some instances, we know Φ, are able to assess Φ, or can estimate Φ using previous data, so that

$$p(\Phi) \quad = \quad \begin{cases} 1, & \text{if } \Phi = \Phi_0 \\ 0, & \text{if } \Phi \neq \Phi_0, \end{cases} \tag{5.2.1}$$

a degenerate distribution. If the observation vectors were independent, then $\Phi_0 = I_N$.

For Gibbs sampling, we will need the conditional posterior distributions. When the covariance matrix Φ is known to be Φ_0, then the only change in posterior conditional distributions for the parameters Ψ, m, F, and Λ is that Φ is now replaced by Φ_0.

Gibbs Sampling Estimation, Φ known

For Gibbs estimation of the posterior, we start with initial values for Ψ, m, F, and Λ say $\bar{\Psi}_{(0)}$, $\bar{m}_{(0)}$, $\bar{F}_{(0)}$, and $\bar{\Lambda}_{(0)}$.

Then for $m = \bar{m}_{(i)}$ cycle through

$$\begin{aligned}
\bar{\Psi}_{(i+1)} &\equiv \text{ a random sample from } p(\Psi|\bar{m}_{(i)}, \bar{F}_{(i)}, \bar{\Lambda}_{(i)}, X) \\
\bar{F}_{(i+1)} &\equiv \text{ a random sample from } p(F|\bar{\Psi}_{(i+1)}, \bar{m}_{(i)}, \bar{\Lambda}_{(i)}, X) \\
\bar{\Lambda}_{(i+1)} &\equiv \text{ a random sample from } p(\Lambda|\bar{\Psi}_{(i+1)}, \bar{m}_{(i)}, \bar{F}_{(i+1)}, X)
\end{aligned}$$

and for $m = \bar{m}_{(i)}$ we have the sequence

$$\begin{gathered}
(\bar{\Psi}_{(1)}, \bar{F}_{(1)}, \bar{\Lambda}_{(1)}) \\
\vdots \\
(\bar{\Psi}_{(s)}, \bar{F}_{(s)}, \bar{\Lambda}_{(s)}) \\
(\bar{\Psi}_{(s+1)}, \bar{F}_{(s+1)}, \bar{\Lambda}_{(s+1)}) \\
\vdots \\
(\bar{\Psi}_{(s+t)}, \bar{F}_{(s+t)}, \bar{\Lambda}_{(s+t)})
\end{gathered}$$

The first s random samples called the "burn in" are discarded and the remaining t samples are kept. The means of the remaining random samples

$$\bar{\Psi} = \frac{1}{t}\sum_{k=1}^{t} \bar{\Psi}_{(s+k)} \tag{5.2.2}$$

$$\bar{F} = \frac{1}{t}\sum_{k=1}^{t} \bar{F}_{(s+k)} \tag{5.2.3}$$

$$\bar{\Lambda} = \frac{1}{t}\sum_{k=1}^{t} \bar{\Lambda}_{(s+k)} \tag{5.2.4}$$

are used as the posterior mean estimates of the parameters given $m = \bar{m}_{(i)}$. We do this for each value of m, then find the value of $m = \bar{m}$ that makes the posterior conditional for the number of factors $p(m|\bar{\Psi}, \bar{F}, \bar{\Lambda}, X)$ a maximum given the corresponding estimates of the other parameters. We will have $(\bar{m}, \bar{\Psi}, \bar{F}, \bar{\Lambda})$ as our posterior estimates of the unknown parameters where $(\bar{\Psi}, \bar{F}, \bar{\Lambda})$ are the estimates conditional on $m = \bar{m}$. It should be noted that LSO is possible when Φ is known because all of the posterior conditional distributions are unimodal so we are sure converge to the global maximum.

5.3 Separable Model, Φ Unknown

In this section, we will consider the case when Φ is an unknown general covariance matrix. When Φ is a general unknown matrix, the number of distinct parameters is enormous so we only outline the procedure and do not carry it out or recommend carrying it out due to its impracticality and computational restrictions.

Prior Distribution

When the covariance matrix between the observations Φ is a full general matrix, we will assume that a priori that it has an inverted Wishart distribution. The distribution for the between observation covariance matrix is given by

$$p(\Phi) \quad = \quad c(N,\eta)|\Phi|^{-\frac{\eta}{2}}e^{-\frac{1}{2}tr\Phi^{-1}D}, \; \Phi, D>0, \; \eta>2N. \qquad (5.3.1)$$

where Φ and D are positive definite matrices and $c(N,\eta)$ is a constant depending only on N and η. Also, D and η are hyperparameters to be assessed.

Posterior Distribution

Using the aforementioned likelihood and prior distributions along with Bayes' Rule and some algebra the joint posterior is given by

$$\begin{aligned} p(\Phi,\Psi,m,F,\Lambda|X) \quad &\propto \quad p(m)|\Phi|^{-\frac{(\eta+p+m)}{2}}e^{-\frac{1}{2}tr\Phi^{-1}V}(2\pi)^{-\frac{(N+p)m}{2}}|H|^{\frac{p}{2}}|\Psi|^{-\frac{(N+m+\nu)}{2}} \\ &\cdot \quad e^{-\frac{1}{2}tr\Psi^{-1}[(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'+B]} \qquad (5.3.2)\end{aligned}$$

where

$$V \equiv (X - F\Lambda')\Psi^{-1}(X - F\Lambda')' + FR^{-1}F' + D. \tag{5.3.3}$$

The Conditional for Φ

From the posterior distribution we can obtain the posterior conditional distribution for Φ.

$$\begin{aligned} p(\Phi|F,\Lambda,\Psi,m,X) &\propto p(F,\Lambda,\Phi,\Psi,m)p(X|F,\Lambda,\Psi,\Phi,m) \\ &\propto p(\Phi)p(\Psi)p(m)p(\Lambda|\Psi,m) \\ &\cdot\ p(F|\Phi,m)p(X|\Psi,\Phi,m,F,\Lambda) \\ &\propto p(\Phi)p(F|\Phi,m)p(X|\Psi,\Phi,m,F,\Lambda) \\ &\propto |\Phi|^{-\frac{(\eta+p+m)}{2}}e^{-\frac{1}{2}tr\Phi^{-1}V} \end{aligned} \tag{5.3.4}$$

where V is as previously defined.

The posterior conditional distribution for the across observation covariance matrix Φ, the within covariance matrix Ψ, the number of factors m, the factor scores F, and the factor loadings is inverted Wishart.

Gibbs Estimation, Φ Unknown

For Gibbs estimation of the posterior, we start with initial values for the unknown parameters Φ, Ψ, m, and F say $\bar{\Phi}_{(0)}$, $\bar{\Psi}_{(0)}$, $\bar{m}_{(0)}$, and $\bar{F}_{(0)}$.

Then for a given number of factors $m = \bar{m}_{(i)}$ cycle through

$$\begin{aligned}
\bar{\Psi}_{(i+1)} &\equiv \text{ a random sample from } p(\Psi|\bar{\Phi}_{(i)}, \bar{F}_{(i)}, \bar{\Lambda}_{(i)}, \bar{m}_{(i)}, X) \\
\bar{F}_{(i+1)} &\equiv \text{ a random sample from } p(F|\bar{\Phi}_{(i)}, \bar{\Psi}_{(i+1)}, \bar{\Lambda}_{(i)}, \bar{m}_{(i)}, X) \\
\bar{\Lambda}_{(i+1)} &\equiv \text{ a random sample from } p(\Lambda|\bar{\Phi}_{(i)}, \bar{\Psi}_{(i+1)}, \bar{F}_{(i+1)}, \bar{m}_{(i)}, X) \\
\bar{\Phi}_{(i+1)} &\equiv \text{ a random sample from } p(\Phi|\bar{\Psi}_{(i+1)}, \bar{F}_{(i+1)}, \bar{\Lambda}_{(i+1)}, \bar{m}_{i}, X)
\end{aligned}$$

and for the given number of factors $m = \bar{m}_{(i)}$ we have the sequence

$$\begin{gathered}
(\bar{\Phi}_{(1)}, \bar{\Psi}_{(1)}, \bar{F}_{(1)}, \bar{\Lambda}_{(1)}) \\
\vdots \\
(\bar{\Phi}_{(s)}, \bar{\Psi}_{(s)}, \bar{F}_{(s)}, \bar{\Lambda}_{(s)}) \\
\vdots \\
(\bar{\Phi}_{(s+t)}, \bar{\Psi}_{(s+t)}, \bar{F}_{(s+t)}, \bar{\Lambda}_{(s+t)})
\end{gathered}$$

The first s random samples called the “burn in” are discarded and the remaining t samples are kept. We use the means of the remaining random samples

$$\bar{F} = \frac{1}{t} \sum_{k=1}^{t} \bar{F}_{(s+k)}$$

$$\bar{\Lambda} = \frac{1}{t} \sum_{k=1}^{t} \bar{\Lambda}_{(s+k)}$$

$$\bar{\Psi} = \frac{1}{t} \sum_{k=1}^{t} \bar{\Psi}_{(s+k)}$$

$$\bar{\Phi} = \frac{1}{t} \sum_{k=1}^{t} \bar{\Phi}_{(s+k)}$$

as the sampling based marginal posterior mean estimates of the parameters given the number of factors $m = \bar{m}_{(i)}$. We do this for each value of the number of factors m, then find the value of the number of factors $m = \bar{m}$ that makes the posterior conditional for the number of factors $p(m|\bar{\Phi}, \bar{\Psi}, \bar{F}, \bar{\Lambda}, X)$ a maximum given the corresponding estimates of the other parameters. This is the same as finding the value for the number of factors that makes the posterior odds ratio a maximum.

We will have $(\bar{m}, \bar{\Phi}, \bar{\Psi}, \bar{F}, \bar{\Lambda})$ as our posterior estimates of the unknown parameters where $(\bar{\Phi}, \bar{\Psi}, \bar{F}, \bar{\Lambda})$ are the estimates conditional on the given value for the number of factors $m = \bar{m}$.

It should be noted that LSO is also possible because all of the posterior conditional distributions are unimodal and we are guaranteed to converge to a global maximum.

5.4 Separable Model, Homoscedastic Structured Φ

It is often the case that Φ is unknown but structured. When Φ is unknown, the conditionals for Ψ, m, F, and Λ do not change from when Φ is known or unknown and general. The structure of the covariance matrix Φ can be determined using covariance determination techniques (see appendix B). Once the structure is determined, we need to add the prior distributions for the unknown parameters in Φ and calculate the posterior conditional distribution for the unknown parameters in Φ. In this section, we will assume that the observations are homoscedastic and consider Φ to be a structured correlation matrix that depends on a single parameter ρ.

There are many possible structures that we are able to specify for Φ that apply to a wide variety of situations. Given that we have homoscedasticity of the observation vectors, then

$$\Omega \equiv \Phi \otimes \Psi = \begin{pmatrix} \Psi & \phi_{12}\Psi & & \cdots & \phi_{1N}\Psi \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \Psi \end{pmatrix}$$

where Φ is a correlation matrix.

We may find that there is a structure in the correlation matrix Φ so that its elements only depend on a single parameter ρ, then the covariance matrix becomes

$$\Omega \equiv \Phi \otimes \Psi = \begin{pmatrix} \Psi & \phi_{12}(\rho)\Psi & & \cdots & \phi_{1N}(\rho)\Psi \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \\ & & & & \Psi \end{pmatrix}.$$

Two well known examples of possible correlation structures for Φ are intraclass and first order Markov. We will state these correlation structures and derive the posterior conditionals for both of these correlations assuming a scaled beta prior distribution.

Φ Represents an Intraclass Correlation

If we determine that the observations are correlated according to an intraclass correlation. An intraclass correlation is used when we have a set of variables and we believe that any two are related in the same way. Any two variables have the same correlation. Then the between observation correlation matrix Φ is

$$\Phi = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ & 1 & \rho & \cdots & \rho \\ & & \ddots & & \vdots \\ & & & & \rho \\ & & & & 1 \end{pmatrix} = (1-\rho)I_N + \rho ee', \qquad (5.4.1)$$

where e is a column vector of ones and $-\frac{1}{N-1} < \rho < 1$ to keep $\Phi > 0$.

Φ Represents a First Order Markov Correlation

We can assume that the observations are correlated according to a first order Markov scheme (Press 1982, p.224). In a first order Markov scheme, we have observations that are related according to a $VAR(1)$. If the subscript i indexes the observations (for example time) then the error structure

$$\epsilon_i = \rho\epsilon_{i-1} + u_i, \tag{5.4.2}$$

where we must restrict ρ by $|\rho| < 1$ in order to keep a constant variance, and u_i is an error term with zero mean and constant variance. With this structure, the between observation correlation matrix Φ is

$$\Phi = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{N-1} & \rho^{N-2} & & \cdots & 1 \end{pmatrix} \tag{5.4.3}$$

where $0 < |\rho| < 1$.

Prior distribution For ρ

If Φ is known except for one parameter as in the cases discussed earlier, then we assess a prior distribution for the unknown parameter ρ, say $p(\rho)$. For example, we could assume that

$$-1 \leq a < \rho < b \leq 1$$

and assess the following scaled beta prior distribution

$$p(\rho) \quad = \quad \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{\rho+a}{b-a}\right)^{\alpha-1} \left(1-\frac{\rho+a}{b-a}\right)^{\beta-1}, \tag{5.4.4}$$

where $\Gamma(\cdot)$ is the gamma function and $\alpha, \beta > 0$ with mean given by

$$E(\rho) \quad = \quad (b-a)\left(\frac{\alpha}{\alpha+\beta}\right) - a. \tag{5.4.5}$$

A scaled beta prior distribution can be used to capture a broad set of beliefs regarding a variable with a fixed range of values. The hyperparameters α and β have the interpretation that $\alpha+\beta-2$ is the effective prior sample size, and a priori, we believe that for every $\alpha-1$ times we believe $\rho = b$ we believe there are $\beta-1$ times $\rho = a$. If for example we expressed no prior beliefs about the value of the parameter ρ then $\alpha = 1$ and $\beta = 1$ can be used which corresponds to a vague or uninformative prior distribution. The assessment of the hyperparameters α and β are discussed in appendix C.

The Conditional for ρ

Using the above mentioned likelihood and associated prior distributions for the parameters of interest, the posterior conditional density for ρ is

$$\begin{aligned}
p(\rho|\Psi, m, F, \Lambda, X) &\propto p(\rho, \Psi, m, F, \Lambda)p(X|\rho, \Psi, m, F, \Lambda) \\
&\propto p(\rho)p(\Psi)p(m)p(F|\Phi, m)p(\Lambda|\Psi, m) \\
&\cdot p(X|\rho, m, F, \Lambda, \Psi) \\
&\propto p(F|\Phi, m)p(\rho)p(X|\rho, m, F, \Lambda, \Psi) \\
&\propto |\Phi|^{-\frac{m}{2}}|R|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Phi^{-1}FR^{-1}F'} \\
&\cdot \left(\frac{\rho + a}{b - a}\right)^{\alpha-1}\left(1 - \frac{\rho + a}{b - a}\right)^{\beta-1} \\
&\cdot |\Phi|^{-\frac{p}{2}}|\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Phi^{-1}(X-F\Lambda')\Psi^{-1}(X-F\Lambda')'} \\
&\propto |\Phi|^{-\frac{(p+m)}{2}}\left(\frac{\rho + a}{b - a}\right)^{\alpha-1}\left(1 - \frac{\rho + a}{b - a}\right)^{\beta-1} e^{-tr\Phi^{-1}C} \quad (5.4.6)
\end{aligned}$$

where

$$C = (X - F\Lambda')\Psi^{-1}(X - F\Lambda')' + FR^{-1}F'. \quad (5.4.7)$$

This is not a well–known recognizable distribution. Random sample generation will be performed by a simple univariate rejection sampling technique.

The exact form of $|\Phi|$ and Φ^{-1} will depend on which structure we determine for Φ.

The Conditional for ρ, Φ Intraclass

As previously stated, the exact form of the conditional posterior distribution depends on which structure we determine for the correlation matrix Φ. If we determine the intraclass structure that has the covariance between any two observations being the same, then we can use the result that the determinant of Φ has the form

$$|\Phi| = (1-\rho)^{N-1}[1+\rho(N-1)] \tag{5.4.8}$$

and the result that the inverse of Φ has the form

$$\Phi^{-1} = \frac{I_N}{1-\rho} - \frac{\rho e e'}{(1-\rho)[1+(N-1)\rho]}, \tag{5.4.9}$$

which is again a matrix with intraclass correlation structure (Press 1982, p.23). Using the aforementioned likelihood, priors, and forms above we obtain

$$\begin{aligned}
p(\rho|F,\Lambda,\Psi,m,X) \;\propto\;& \left(\frac{\rho+a}{b-a}\right)^{\alpha-1}\left(1-\frac{\rho+a}{b-a}\right)^{\beta-1} \\
\cdot\;& |\Phi|^{-\frac{(p+m)}{2}} e^{-\frac{1}{2}tr\Phi^{-1}C} \\
\propto\;& \left(\frac{\rho+a}{b-a}\right)^{\alpha-1}\left(1-\frac{\rho+a}{b-a}\right)^{\beta-1} \\
\cdot\;& (1-\rho)^{-\frac{(N-1)(p+m)}{2}}[1+\rho(N-1)]^{-\frac{(p+m)}{2}} \\
\cdot\;& e^{-\frac{1}{2}\left(\frac{1}{1-\rho}trI_NC-\frac{\rho}{(1-\rho)[1+(N-1)\rho]}tree'C\right)} \\
\propto\;& \left(\frac{\rho+a}{b-a}\right)^{\alpha-1}\left(1-\frac{\rho+a}{b-a}\right)^{\beta-1} \\
\cdot\;& (1-\rho)^{-\frac{(N-1)(p+m)}{2}}[1+\rho(N-1)]^{-\frac{(p+m)}{2}} \\
\cdot\;& e^{-\frac{1}{2}\left(\frac{c_1}{1-\rho}-\frac{c_2\rho}{(1-\rho)[1+(N-1)\rho]}\right)}
\end{aligned} \tag{5.4.10}$$

where C is as defined as before,

$$c_1 = tr(C),$$

and

$$c_2 = tr(ee'C).$$

This is not recognizable as a friendly distribution so we must use use rejection sampling in order to generate random samples.

The Conditional for ρ, Φ First Order Markov

If we determine the first order Markov structure, then we can use the result that the determinant of a matrix with such structure has the form

$$|\Phi| = (1-\rho^2)^{N-1} \tag{5.4.11}$$

and the result (Press 1982, p.24) that the inverse of such a patterned matrix has the form

$$\Phi^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & & & 0 \\ -\rho & (1+\rho^2) & -\rho & & \\ & \ddots & \ddots & \ddots & \\ & & & (1+\rho^2) & -\rho \\ 0 & & & -\rho & 1 \end{pmatrix}. \tag{5.4.12}$$

along with the aforementioned likelihood and prior distributions to obtain

$$
\begin{aligned}
p(\rho|F,\Lambda,\Psi,m,X) &\propto \left(\frac{\rho+a}{b-a}\right)^{\alpha-1}\left(1-\frac{\rho+a}{b-a}\right)^{\beta-1} \\
&\cdot \quad |\Phi|^{-\frac{(p+m)}{2}} e^{-\frac{1}{2}tr\Phi^{-1}C} \\
&\propto \left(\frac{\rho+a}{b-a}\right)^{\alpha-1}\left(1-\frac{\rho+a}{b-a}\right)^{\beta-1} \\
&\cdot \quad (1-\rho^2)^{-\frac{(N-1)(p+m)}{2}} e^{-\frac{1}{2}tr(M_1-\rho M_2+\rho^2 M_3)C/(1-\rho^2)} \\
&\propto \left(\frac{\rho+a}{b-a}\right)^{\alpha-1}\left(1-\frac{\rho+a}{b-a}\right)^{\beta-1} \\
&\cdot \quad (1-\rho^2)^{-\frac{(N-1)(p+m)}{2}} e^{-\frac{1}{2}(k_1-k_2\rho+k_3\rho^2)/(1-\rho^2)}
\end{aligned}
$$

where the matrices and constants used are, C as defined previously,

$$M_1 = I_N,$$

$$M_2 = \begin{pmatrix} 0 & 1 & & & 0 \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 0 & 1 \\ 0 & & & 1 & 0 \end{pmatrix},$$

$$M_3 = \begin{pmatrix} 0 & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ 0 & & & & 0 \end{pmatrix},$$

$$k_1 = tr(C),$$

$$k_2 = tr(M_2 C),$$

and

$$k_3 = tr(M_3C).$$

Again, this is not recognizable as a friendly distribution. We must use rejection sampling in order to generate random samples from this conditional posterior distribution.

These are two simple possible structures. There may be others that also depend on a single parameter or on several parameters. For the purpose of illustration, we will only study the two aforementioned structures. The rejection sampling technique is simple to carry out because we only need to generate samples from a univariate distribution.

Gibbs Estimation, Φ Unknown but Structured

For Gibbs estimation of the posterior, we start with initial values for the parameters ρ, Ψ, m, and F say $\bar{\rho}_{(0)}$, $\bar{\Psi}_{(0)}$, $\bar{m}_{(0)}$, and $\bar{F}_{(0)}$.

Then for a given number of factors $m = \bar{m}_{(i)}$ cycle through

$$\begin{aligned}
\bar{\Psi}_{(i+1)} &\equiv \text{ a random sample from } p(\Psi|\bar{\rho}_{(i)}, \bar{F}_{(i)}, \bar{\Lambda}_{(i)}, \bar{m}_{(i)}, X) \\
\bar{F}_{(i+1)} &\equiv \text{ a random sample from } p(F|\bar{\rho}_{(i)}, \bar{\Psi}_{(i+1)}, \bar{\Lambda}_{(i)}, \bar{m}_{(i)}, X) \\
\bar{\Lambda}_{(i+1)} &\equiv \text{ a random sample from } p(\Lambda|\bar{\rho}_{(i)}, \bar{\Psi}_{(i+1)}, \bar{F}_{(i+1)}, \bar{m}_{(i)}, X) \\
\bar{\rho}_{(i+1)} &\equiv \text{ a random sample from } p(\rho|\bar{\Psi}_{(i+1)}, \bar{F}_{(i+1)}, \bar{\Lambda}_{(i+1)}, \bar{m}_{(i)}, X)
\end{aligned}$$

and for the given value for the number of factors $m = \bar{m}_{(i)}$ we have the sequence

$$\begin{gathered}
(\bar{\rho}_{(1)}, \bar{\Psi}_{(1)}, \bar{F}_{(1)}, \bar{\Lambda}_{(1)}) \\
\vdots \\
(\bar{\rho}_{(s)}, \bar{\Psi}_{(s)}, \bar{F}_{(s)}, \bar{\Lambda}_{(s)}) \\
(\bar{\rho}_{(s+1)}, \bar{\Psi}_{(s+1)}, \bar{F}_{(s+1)}, \bar{\Lambda}_{(s+1)}) \\
\vdots \\
(\bar{\rho}_{(s+t)}, \bar{\Psi}_{(s+t)}, \bar{F}_{(s+t)}, \bar{\Lambda}_{(s+t)}).
\end{gathered}$$

The first s random samples called the "burn in" are discarded and the remaining t samples are kept to be used for our estimates. We use the means of the remaining t random samples

$$\bar{F} = \frac{1}{t} \sum_{k=1}^{t} \bar{F}_{(s+k)}$$

$$\bar{\Lambda} = \frac{1}{t} \sum_{k=1}^{t} \bar{\Lambda}_{(s+k)}$$

$$\bar{\Psi} = \frac{1}{t} \sum_{k=1}^{t} \bar{\Psi}_{(s+k)}$$

$$\bar{\rho} = \frac{1}{t} \sum_{k=1}^{t} \bar{\rho}_{(s+k)}$$

as the sampling based posterior mean estimates of the parameters for a given number of factors $m = \bar{m}_{(i)}$.

We carry out this procedure of sampling and calculating the means of the remaining samples for each value of the number of factors m, then find the value of of the number of factors $m = \bar{m}$ that makes the posterior conditional distribution for the number of factors $p(m|\bar{\rho}, \bar{\Psi}, \bar{F}, \bar{\Lambda}, X)$ a maximum given the corresponding estimates of the other parameters. This is the same as selecting the number of factors to be that value which makes the conditional posterior odds ratio a maximum. We will have $(\bar{m}, \bar{\rho}, \bar{\Psi}, \bar{F}, \bar{\Lambda})$ as our posterior estimates of the unknown parameters where $(\bar{\rho}, \bar{\Psi}, \bar{F}, \bar{\Lambda})$ are the estimates conditional on $m = \bar{m}$. The step where we draw samples from $p(\rho|\Psi, m, F, \Lambda, X)$ is performed by univariate rejection sampling.

6 Matrix Intraclass Covariance Model

When the observation vectors have a matrix intraclass covariance matrix ($\Omega_{ii} = \Psi$ and $\Omega_{ij} = \Upsilon$, $i \neq j$.), we use the following [1]. Note that if $\Upsilon = 0$, the null matrix, then we have independent observations and if $\Upsilon = \rho\Psi$ we have the separable intraclass covariance discussed earlier.

6.1 Likelihood

The likelihood function for the observations when the errors are assumed to have a matrix intraclass structure (after centering them about the sample mean) is given by

$$p(x|\Psi, \Upsilon, m, f, \Lambda) = (2\pi)^{-\frac{Np}{2}} |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2}[x-(I_N \otimes \Lambda)f]'\Omega^{-1}[x-(I_N \otimes \Lambda)f]}. \tag{6.1.1}$$

where the covariance matrix is

$$\Omega = \left(\begin{array}{ccccc} \Psi & \Upsilon & & \cdots & \Upsilon \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \Upsilon \\ & & & & \Psi \end{array} \right), \tag{6.1.2}$$

where $\Psi > 0$, $\Upsilon > 0$ and Ψ is assumed to be diagonal on average. We will make an orthogonal transformation on our observations. Let $y = \Gamma x = (y_1', \ldots, y_N')'$, with Γ

[1] Thanks to Dr. S. James Press for suggesting this covariance structure and the transformation of the observations that results in a new covariance matrix that is block diagonal.

such that $\Gamma\Gamma' = I$, $\Gamma = \Gamma_0 \otimes I_p$, Γ_0 a Helmert matrix, then the likelihood for the transformed variables is given by

$$(y|\Psi, \Upsilon, m, f, \Lambda) \sim N\left(\Gamma(I_N \otimes \Lambda)f, \Gamma\Omega\Gamma'\right) \tag{6.1.3}$$

which after some algebra and using Theorem (5) of (Press 1979) becomes

$$(y|\chi, \Xi, m, f^*, \Lambda) \sim N\left((I_N \otimes \Lambda)f^*, D_1\right) \tag{6.1.4}$$

where the transformed factor scores f^* is

$$f^* = (\Gamma_0 \otimes I_m)f, \tag{6.1.5}$$

and the covariance matrix for the transformed observations is

$$D_1 = \begin{pmatrix} \chi & & & & \\ & \Xi & & 0 & \\ & & \ddots & & \\ & 0 & & & \\ & & & & \Xi \end{pmatrix}, \quad D_1 > 0. \tag{6.1.6}$$

The transformed observations are independent with covariance matrices

$$\chi = \Psi + (N-1)\Upsilon, \quad \chi > 0 \tag{6.1.7}$$

for the first transformed variable and

$$\Xi = \Psi - \Upsilon, \quad \Xi > 0. \tag{6.1.8}$$

for the remaining $N-1$ transformed variables.

We can partition y and f^* into

$$y = \begin{pmatrix} y_1 \\ z \end{pmatrix}, \quad f^* = \begin{pmatrix} f_1^* \\ g \end{pmatrix}, \tag{6.1.9}$$

where y_1 and f_1^* are $p \times 1$ vectors while z and g are $(N-1)p \times 1$ vectors. It is readily seen from (6.1.4-6.1.8) that

$$\left.\begin{array}{l} (y_1|\chi, m, f_1^*, \Lambda) \sim N(\Lambda f_1^*, \chi) \\ (z|\Xi, m, g, \Lambda) \sim N\left((I_{N-1} \otimes \Lambda)g, I_{N-1} \otimes \Xi\right) \end{array}\right\} independent \tag{6.1.10}$$

or the likelihood for the first transformed observation is

$$p(y_1|\chi, m, f_1^*, \Lambda) = (2\pi)^{-\frac{p}{2}} |\chi|^{-\frac{1}{2}} e^{-\frac{1}{2}(y_1 - \Lambda f_1^*)'\chi^{-1}(y_1 - \Lambda f_1^*)}. \tag{6.1.11}$$

and the likelihood for the remaining $N-1$ transformed observations is

$$p(z|\Xi, m, g, \Lambda) = (2\pi)^{-\frac{(N-1)p}{2}} |I_{N-1} \otimes \Xi|^{-\frac{1}{2}} e^{-\frac{1}{2}[z - (I_{N-1} \otimes \Lambda)g]'(I_{N-1} \otimes \Xi)^{-1}[z - (I_{N-1} \otimes \Lambda)g]}. \tag{6.1.12}$$

The likelihood of the last $N-1$ transformed observations may be rewritten as

$$p(Z|\Xi, m, G, \Lambda) = (2\pi)^{-\frac{np}{2}}|\Xi|^{-\frac{n}{2}}e^{-\frac{1}{2}tr\Xi^{-1}(Z-G\Lambda')'(Z-G\Lambda')} \tag{6.1.13}$$

where $Z' = (z_1, \ldots, z_n)$, $G' = (g_1, \ldots, g_n)$, $n = N-1$, and Ξ is assumed to be diagonal on average.

We will neglect the first transformed observation thus we have n independent transformed observations.

6.2 Prior Distributions

We will use generalized natural conjugate families of prior distributions for the parameters. Lets consider the prior distribution for the factor scores

$$p(f|R, P, m) \quad = \quad (2\pi)^{-\frac{Nm}{2}}|\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}f'\Theta^{-1}f}. \tag{6.2.1}$$

where we assume the matrix intraclass correlation matrix

$$\Theta = \begin{pmatrix} R & P & & \cdots & P \\ & R & & & \\ & & \ddots & & \vdots \\ & & & & P \\ & & & & R \end{pmatrix}, \quad R > 0, \; P > 0.$$

In performing the transformation given above from f to f^*, the prior distribution for the factor scores becomes

$$p(f^*|R_1, R_2, m) = (2\pi)^{-\frac{Nm}{2}} |D_2|^{-\frac{1}{2}} e^{-\frac{1}{2}f' D_2^{-1} f} \tag{6.2.2}$$

and the covariance matrix for the transformed factor scores is

$$D_2 = \begin{pmatrix} R_1 & & & \\ & R_2 & & 0 \\ & & \ddots & \\ & 0 & & \\ & & & R_2 \end{pmatrix}, \quad D_2 > 0. \tag{6.2.3}$$

The transformed factor score vectors are independent with covariance matrices

$$R_1 = R + (N-1)P, \quad R_1 > 0 \tag{6.2.4}$$

for the first transformed factor score vector and

$$R_2 = R - P, \quad R_2 > 0. \tag{6.2.5}$$

for the remaining $N - 1$ transformed factor score vectors.

Using the partition $f^* = (f_1^*, g)$

$$\left.\begin{array}{l} (f_1^*|R_1, m) \sim N(0, R_1) \\ (g|R_2, m) \sim N\left(0, I_{N-1} \otimes R_2\right) \end{array}\right\} independent \tag{6.2.6}$$

or the prior for the first transformed observation is

$$p(f_1^*|R_1,m) \quad = \quad (2\pi)^{-\frac{p}{2}}|R_1|^{-\frac{1}{2}}e^{-\frac{1}{2}f_1^{*\prime}R_1^{-1}f_1^*} \tag{6.2.7}$$

and the prior distribution for the remaining $N-1$ transformed factor scores is

$$p(G|R_2,m) \quad = \quad (2\pi)^{-\frac{nm}{2}}|R_2|^{-\frac{n}{2}}e^{-\frac{1}{2}trGR_2^{-1}G'}. \tag{6.2.8}$$

We will neglect the first transformed factor score vector.

Returning to the prior distributions. We will assume that the joint prior distribution for the parameters in the likelihood containing n transformed observation vectors is

$$p(\Xi,G,\Lambda,m) \quad = \quad p(\Xi)p(m)p(G|R_2,m)p(\Lambda|\Xi,m). \tag{6.2.9}$$

where

$$\begin{aligned}
p(\Xi) &= c(n,\nu)|\Xi|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Xi^{-1}B},\ \Xi>0,\ \nu>2n, && (6.2.10)\\
p(m) &= \text{a discrete distribution to be defined below} && (6.2.11)\\
p(G|R_2,m) &= (2\pi)^{-\frac{nm}{2}}|R_2|^{-\frac{n}{2}}e^{-\frac{1}{2}trGR_2^{-1}G'},\quad R_2>0 && (6.2.12)\\
p(\Lambda|\Xi,m) &= (2\pi)^{-\frac{pm}{2}}|\Xi|^{-\frac{m}{2}}e^{-\frac{1}{2}(\Lambda-\Lambda_0)\Xi^{-1}(\Lambda-\Lambda_0)'},\quad \Xi>0, && (6.2.13)
\end{aligned}$$

6.3 Posterior Distribution

By Bayes' rule and some algebra, the posterior distribution of the unknown parameters becomes

$$\begin{aligned} p(\Xi, m, G, \Lambda | Z) &\propto p(m)p(\Xi)(2\pi)^{-\frac{(n+p)m}{2}}|H|^{\frac{p}{2}}|\Xi|^{-\frac{(n+m+\nu)}{2}} \\ &\cdot \quad |R_2|^{-\frac{n}{2}}e^{-\frac{1}{2}trGR_2^{-1}G'}e^{-\frac{1}{2}tr\Xi^{-1}U} \end{aligned} \qquad (6.3.1)$$

where the posterior conditional mean is given by

$$U \equiv (Z - G\Lambda')'(Z - G\Lambda') + (\Lambda - \Lambda_0)H(\Lambda - \Lambda_0)' + B. \qquad (6.3.2)$$

6.4 Estimation

We simply note that N, Ψ, and F, are replaced by n, Ξ, and G in the priors, posterior, the conditionals, and the estimation algorithm of the independent observation model. We may now estimate the parameters by either conditional modal, LSO, or Gibbs estimation.

7 Examples

We present two examples in this section. The first is a simulation to compare CBFA and PS89 to ground truth. The second example, is an analysis of a real data set.

7.1 Simulation Example

For our first example, we simulated a set of data to compare CBFA and PS89 in a known setting. We determined the number of factors using the correlated Bayesian factor analysis (CBFA) methods previously developed and compared the results to those of the PS89 model.

Data

We are using the homoscedastic separable covariance model that depends on a single parameter ρ. For the correlation matrix Φ, we have selected the first order Markov autocorrelation scheme. The data we simulated was of size $N = 100$ with dimension $p = 12$. We use the convention of denoting the true known parameter by using an asterisk as a subscript.

We selected the true correlation parameter to be $\rho_* = 0.25$ and the disturbance covariance matrix to be $\Psi_* = 25I_p$. We then selected the true number of factors to be $m_* = 4$ and the true factor loading matrix to be as given in the estimation section.

We selected the true correlation matrix for the factor scores to be $R = I_M$

corresponding to there being independence within the factor score vectors. We generated a random score f_* from

$$N(0, \Phi_* \otimes I_m)$$

and a random error ϵ_* from

$$N(0, \Phi_* \otimes \Psi_*)$$

to obtain our simulated data

$$X_* = F_* \Lambda_*' + E_*$$

where $f_* = vec(F_*')$ and $\epsilon_* = vec(E_*')$.

We partitioned the generated data into $(Y, X)'$. The first half for parameter estimation and the second half for analysis.

Assessment

We took the first 50 observations Y and use these to assess our hyperparameters then analyzed the remaining 50 observations X with the assessed hyperparameters.

We selected the previously stated prior distributions for the separable model (equations 5.1.5 - 5.1.7 and 5.4.4).

The mean of the of the training data used for hyperparameter assessment is given in Table 3 while the covariance matrix is given in Table 4.

Table 3: Simulation Training Mean.

p	mean
1	-3.902
2	-2.101
3	-2.306
4	-0.723
5	-2.828
6	-4.032
7	2.742
8	2.870
9	2.633
10	1.007
11	1.811
12	2.457

The hyperparameters were assessed according to method (a) in appendix C. The hyperparameters are $\nu = 76$ and $b_0 = 1105.9$. The possible number of factors $m = 3$, 4, 5 was determined from a principal components analysis on the covariance matrix (Table 4) of the training data because they had eigenvalues of 2.574, 1.585, and 0.373 respectively. These factors also accountes for 74.8%, 88.0%, 91.1% of the variance respectively. The prior $p(3) = p(4) = p(5) = \frac{1}{3}$ was assessed.

We assessed the a priori mean for the factor loadings to be as displayed in the next section, the prior matrix H to be $H = \frac{1}{50} I_m$, the prior correlation matrix for the factor scores to be $R = I_m$ because we wish to fit the standard orthogonal factor analysis model.

The hyperparameters α and β have the interpretation that $\alpha + \beta - 2$ is the effective prior sample size, and a priori, we believe that for every $\alpha - 1$ times we

Table 4: Simulation Training Covariance matrix.

	1	2	3	4	5	6	7	8	9	10	11	12
1	155.0	145.6	134.5	-8.0	-3.9	-2.8	-40.0	-39.0	-43.4	1.9	0.1	16.8
2		177.2	140.0	-13.5	1.5	-13.4	-52.4	-51.1	-50.5	-1.4	-5.3	6.4
3			164.4	-35.0	-29.8	-28.6	-37.7	-43.9	-52.7	7.6	13.0	18.5
4				123.3	94.7	116.8	-0.9	0.3	4.5	-8.4	11.7	14.5
5					107.6	97.0	-7.1	-1.1	6.6	-13.6	-9.8	4.1
6						139.6	-10.6	-11.9	-4.4	-5.4	17.0	24.7
7							131.3	112.2	110.7	26.3	11.7	16.9
8								132.2	103.8	29.2	2.7	6.2
9									137.4	15.6	-7.7	-0.2
10										101.3	67.9	63.5
11											92.0	68.1
12												88.8

believe $\rho = b$ we believe there are $\beta - 1$ times $\rho = a$. We further selected the range of values for the correlation parameter ρ to be $a = 0$ and $b = \frac{1}{2}$ while we selected the hyperparameter for its prior distribution to be $\alpha = 10$, and $\beta = 10$.

Estimates

Implementing the Gibbs sampling was a computational challenge. Exact implementation required the Cholesky factorization of a large matrix. It is $\Phi \otimes (I_m + \Lambda'\Psi^{-1}\Lambda)^{-1}$ which could not be implemented within a reasonable amount of time using the FORTRAN IMSL library subroutines for exact Cholesky factorization of matrices. Instead of computing the exact Cholesky factorization, we make an approximation in order to reduce computational time. We assumed that $\rho^5 = 0$. The covariance matrix across the factor vectors which is of the form

$$\Theta \equiv \begin{pmatrix} \tilde{R} & \rho\tilde{R} & \rho^2\tilde{R} & \cdots & \rho^{N-1}\tilde{R} \\ & \tilde{R} & & & \\ & & & \ddots & \vdots \\ & & & & \tilde{R} \end{pmatrix},$$

became the band symmetric matrix

$$\Theta \equiv \begin{pmatrix} \tilde{R} & \dots & \rho^4\tilde{R} & & 0 \\ \vdots & \tilde{R} & & \ddots & \\ \rho^4\tilde{R} & \ddots & \ddots & & \\ & & & & \\ 0 & & & & \end{pmatrix},$$

where the matrix $\tilde{R} = (I_m + \Lambda'\Psi^{-1}\Lambda)^{-1}$. That is, we have a band consisting of five blocks. We then used the FORTRAN subroutine SPBTRF for exact Cholesky factorization of band symmetric matrices which is part of the LAPACK library of routines. Computation time[2] on a Sun Ultra 10 for 110000 iterations including 10000 for a burn in is given in Table 5.

Number of Factors	$m = 3$	$m = 4$	$m = 5$
Computation Time	2 hrs 33	3 hrs 2 min	3 hrs 13

Table 5: Simulation Computation time for 110000 iterations.

Upon implementing the CBFA model, we found the number of factors to be $\bar{m} = 4$ as evidenced in the following table containing the log of the conditional posterior probabilities (with an additive constant).

Number of Factors	$m = 3$	$m = 4$	$m = 5$
Log of Posterior	-4745.5	-3956.2	-4493.9

Table 6: Simulation Posterior distribution for the number of factors.

Thus, CBFA correctly determined the number of factors to be four.

We found the correlation parameter to be as given in the Table 7 which is consistent with its true value. We found the estimated factor scores from both

[2] We implemented several parallel runs with 25000 iterations (5000 for burn in) and found the results to be consistent. The smaller runs took approximately 40 minutes for four factors.

Model	Estimate of ρ
True	0.25
CBFA, Φ Markov	0.2501

Table 7: Simulation Estimate of the Correlation Parameter ρ.

CBFA and PS89 to be as given in Table 8.

Table 8: Simulation Factor Scores

	True Factor Scores, F_*.				PS89 Factor Scores, $\hat{F}$.				CBFA Factor Scores, $\bar{F}$.			
N	1	2	3	4	1	2	3	4	1	2	3	4
1	-0.066	-1.268	0.402	0.354	-0.096	-0.926	0.184	0.433	-0.087	-0.916	0.201	0.411
2	-0.219	-0.082	0.417	0.250	0.638	-0.053	0.635	0.123	0.638	-0.058	0.622	0.122
3	-0.281	1.258	-0.153	-0.548	0.037	1.196	-0.187	-0.594	0.019	1.190	-0.212	-0.563
4	-0.577	0.419	-1.023	0.516	-0.224	0.776	-0.920	0.399	-0.228	0.801	-0.948	0.443
5	-0.989	-0.653	0.186	0.517	-0.258	-0.782	0.146	0.625	-0.256	-0.780	0.160	0.603
6	1.956	-0.150	-0.774	1.494	1.371	0.075	-0.447	1.065	1.386	0.083	-0.487	1.094
7	0.656	-0.704	-0.569	-1.475	0.653	-0.559	0.140	-1.427	0.670	-0.577	0.163	-1.455
8	-0.613	-2.516	-1.348	0.154	-0.760	-2.206	-1.039	0.207	-0.725	-2.183	-0.993	0.175
9	0.573	-1.939	-0.812	0.340	0.335	-1.471	-0.605	0.249	0.369	-1.466	-0.573	0.215
10	1.684	-0.827	0.126	0.139	1.771	-0.768	0.119	0.063	1.797	-0.779	0.116	0.042
11	1.222	0.530	0.803	-0.352	1.740	0.509	1.051	-0.546	1.735	0.472	1.025	-0.546
12	-0.754	0.684	-2.550	-0.585	-0.575	0.373	-2.582	-1.296	-0.556	0.394	-2.585	-1.244
13	-1.403	0.327	-0.715	0.030	-1.022	-0.224	-0.524	-0.068	-1.023	-0.213	-0.511	-0.063
14	1.821	-0.226	0.392	0.652	1.915	0.434	1.555	0.272	1.913	0.405	1.522	0.262
15	-0.950	1.919	2.254	1.246	-0.042	2.218	2.584	1.503	-0.104	2.203	2.521	1.537
16	1.093	-1.160	1.195	1.091	1.063	-1.083	1.384	0.719	1.071	-1.100	1.387	0.690
17	-0.744	0.015	0.275	-0.015	-0.120	-0.063	0.082	0.073	-0.116	-0.062	0.092	0.074
18	-0.074	1.172	-0.148	0.126	0.073	1.208	-0.237	0.007	0.060	1.215	-0.257	0.034
19	0.379	-0.622	0.934	-0.208	1.088	-0.425	0.962	-0.287	1.093	-0.449	0.979	-0.327
20	-2.053	-2.607	-0.267	0.584	-1.313	-1.987	-0.286	0.673	-1.299	-1.965	-0.235	0.633
21	-0.782	0.848	-0.492	-2.031	-0.798	0.797	-0.346	-2.223	-0.811	0.775	-0.333	-2.213
22	-0.737	0.685	-0.381	-1.920	-0.678	0.785	-0.221	-2.454	-0.686	0.768	-0.213	-2.440
23	-0.657	1.247	-2.587	-0.136	-0.466	0.893	-1.946	-0.205	-0.452	0.924	-1.958	-0.153
24	1.078	0.106	-1.778	0.104	1.309	0.183	-0.603	-0.775	1.337	0.186	-0.612	-0.758
25	1.402	2.181	0.210	-0.701	1.656	1.990	0.635	-1.537	1.646	1.961	0.592	-1.503
26	0.018	3.027	0.096	-1.476	0.502	3.032	-0.048	-1.394	0.467	3.007	-0.101	-1.338
27	-2.013	1.934	1.852	0.573	-0.730	2.235	1.156	0.766	-0.788	2.224	1.120	0.795
28	-0.647	0.666	2.083	0.080	0.244	0.666	2.009	0.447	0.212	0.640	1.993	0.432
29	-1.881	0.525	0.403	0.052	-1.227	0.619	0.158	-0.051	-1.245	0.630	0.159	-0.049
30	0.713	-0.675	0.213	-0.597	0.923	-0.601	0.351	-0.990	0.935	-0.622	0.363	-1.017
31	-1.129	-0.283	0.845	-0.997	-0.678	-0.266	-0.371	-1.067	-0.682	-0.274	-0.340	-1.089
32	-1.981	0.519	0.816	-0.110	-1.399	0.169	0.748	-0.324	-1.425	0.170	0.772	-0.349
33	-0.714	-0.515	-0.138	1.009	-0.380	-0.421	-0.522	0.820	-0.371	-0.397	-0.505	0.804
34	-0.693	-1.366	-2.149	0.670	-0.807	-1.605	-1.997	0.109	-0.760	-1.562	-1.943	0.097
35	-0.719	-0.769	-0.921	0.047	-0.397	-0.886	-1.262	-0.398	-0.371	-0.875	-1.225	-0.412
36	-0.577	-1.461	-1.114	-0.269	-0.490	-0.913	-0.980	-0.158	-0.464	-0.886	-0.956	-0.158
37	-0.486	-0.728	-2.100	0.930	-0.463	-0.450	-1.930	0.222	-0.429	-0.410	-1.918	0.240
38	-1.856	0.229	-0.801	-0.904	-1.412	-0.055	-0.610	-1.413	-1.416	-0.048	-0.588	-1.412
39	-0.591	-0.543	1.066	-0.347	-0.266	-0.249	0.869	0.028	-0.281	-0.259	0.861	0.031
40	-2.071	-0.410	1.635	0.179	-1.518	-0.837	1.581	0.326	-1.537	-0.848	1.633	0.267
41	-0.527	-2.246	0.873	1.115	-0.348	-2.149	0.590	1.138	-0.339	-2.150	0.632	1.072
42	-0.278	-1.538	-0.028	2.073	-0.462	-1.129	-0.142	1.956	-0.454	-1.106	-0.131	1.938
43	-0.701	-0.808	1.327	0.801	-0.216	-0.001	1.054	0.156	-0.236	-0.013	1.055	0.137
44	-0.160	-0.120	1.573	3.746	0.344	-0.009	1.370	4.220	0.316	0.010	1.314	4.229
45	-0.052	0.018	0.129	1.281	0.346	0.034	-0.492	0.970	0.362	0.053	-0.501	0.978
46	0.872	2.488	-0.444	0.869	1.363	2.374	0.042	0.325	1.344	2.367	-0.044	0.401
47	0.055	-0.997	-1.984	0.067	-0.352	-1.075	-1.772	0.182	-0.324	-1.047	-1.765	0.205
48	-0.365	0.189	-0.238	-0.710	-0.245	0.059	-0.236	-0.672	-0.248	0.047	-0.232	-0.670
49	0.030	-0.122	-1.180	-0.447	-0.065	0.157	-0.847	-0.548	-0.063	0.149	-0.846	-0.541
50	-0.051	0.082	1.341	0.718	0.434	0.410	1.747	0.351	0.404	0.378	1.731	0.333

Here are the true factor loadings along with the prior and estimated factor loadings. We find that the CBFA estimates look more like the true values than the PS89 estimates.

Table 9: Simulation Factor Loadings

True Loadings, Λ_*.

p	1	2	3	4
1	10	0	0	0
2	10	0	0	0
3	10	10	0	0
4	0	10	0	0
5	0	10	0	0
6	0	0	10	0
7	0	0	10	0
8	0	0	10	0
9	0	0	0	0
10	0	0	0	10
11	0	0	0	10
12	0	0	0	10

Prior Loadings, Λ_0.

p	1	2	3	4
1	11.57	0.18	-1.80	0.45
2	12.12	-0.17	-2.62	-0.40
3	11.14	-2.46	-2.42	1.46
4	-0.91	10.40	0.08	0.59
5	-0.16	9.21	0.20	-1.08
6	-0.78	10.87	-0.93	1.41
7	-1.81	-0.53	10.44	1.62
8	-1.93	-0.38	10.12	0.92
9	-2.17	0.33	10.13	-0.36
10	0.18	-1.08	1.79	8.05
11	-0.22	0.33	-0.42	8.56
12	1.02	1.25	0.41	7.94

PS89 Loadings, $\hat{\Lambda}$.

p	1	2	3	4
1	11.86	-0.56	-1.70	0.49
2	12.05	-0.41	-2.51	0.07
3	11.77	-1.27	-2.21	0.85
4	-0.78	10.56	-0.20	0.37
5	0.29	10.15	0.10	-0.79
6	-1.02	11.42	-0.45	0.99
7	-1.52	-0.36	10.55	1.26
8	-1.59	0.21	10.86	0.74
9	-2.51	-0.27	10.69	0.25
10	0.17	-0.65	1.10	8.42
11	-0.42	-0.47	0.41	9.22
12	1.11	1.17	0.25	8.64

CBFA Loadings, $\bar{\Lambda}$.

p	1	2	3	4
1	11.60	-0.35	-1.64	0.40
2	11.71	-0.46	-2.39	0.04
3	11.50	-1.28	-2.14	0.98
4	-0.79	10.28	-0.08	0.27
5	0.27	9.92	0.18	-0.74
6	-0.92	11.18	-0.36	0.91
7	-1.60	-0.38	10.36	1.35
8	-1.48	0.39	10.55	0.68
9	-2.31	-0.07	10.47	0.40
10	0.14	-0.82	1.21	8.34
11	-0.32	-0.63	0.52	9.78
12	1.02	1.19	0.32	8.29

Table 10: Simulation Disturbance Covariance Matrices

True Disturbance Covariance matrix, Ψ_*.

	1	2	3	4	5	6	7	8	9	10	11	12
1	25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3			25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4				25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5					25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6						25.00	0.00	0.00	0.00	0.00	0.00	0.00
7							25.00	0.00	0.00	0.00	0.00	0.00
8								25.00	0.00	0.00	0.00	0.00
9									25.00	0.00	0.00	0.00
10										25.00	0.00	0.00
11											25.00	0.00
12												25.00

PS89 Disturbance Covariance Matrix, $\hat{\Psi}_{mode}$.

	1	2	3	4	5	6	7	8	9	10	11	12
1	13.03	-1.58	-2.80	-0.07	-0.36	-1.02	-0.62	-1.29	1.66	0.18	0.77	-0.32
2		12.79	-3.10	-0.73	-0.96	0.67	-0.08	-0.19	0.46	0.39	0.30	0.51
3			15.54	0.95	1.39	0.65	1.04	2.08	-2.42	-0.96	-0.75	-0.72
4				14.71	-1.24	-4.48	0.53	0.40	-1.63	-0.03	-0.75	0.33
5					14.26	-2.54	-1.48	0.25	0.47	1.99	-0.74	0.42
6						15.75	1.56	0.20	-0.30	-0.72	0.21	-1.61
7							16.45	-2.81	-4.77	-1.21	-0.44	-0.87
8								14.92	-2.22	-2.34	1.07	0.40
9									15.70	1.53	1.01	0.19
10										15.41	-3.32	-1.84
11											13.64	-1.41
12												13.28

CBFA Disturbance Covariance Matrix, $\bar{\Psi}$.

	1	2	3	4	5	6	7	8	9	10	11	12
1	16.62	0.11	-0.69	0.29	-0.33	-1.01	-0.24	-1.47	1.63	0.09	0.85	-0.28
2		16.41	-1.03	-0.85	-0.42	0.31	-0.17	0.35	-0.15	0.23	0.16	0.76
3			18.98	0.43	1.39	1.26	0.82	2.35	-1.77	-0.94	-0.73	-0.68
4				17.31	0.85	-2.02	0.46	0.42	-1.30	-0.85	-0.37	0.65
5					17.20	-0.39	-1.90	0.93	1.25	1.87	-1.42	0.22
6						19.62	2.02	0.46	-1.09	-0.78	-0.01	-1.28
7							19.91	-0.44	-3.14	-0.93	-0.34	-0.99
8								18.18	0.25	-2.54	0.78	0.70
9									19.01	2.00	0.95	0.15
10										18.94	-1.28	-0.17
11											17.77	0.54
12												16.83

Comparison of CBFA and PS89 Estimators.

So we will address the question of whether or not CBFA is worth the extra work. The purpose of factor analysis is to represent the observations in terms of a smaller set of variables called factors along with factor loadings. The factor scores and loadings capture the essence of the observations by making use of interdependencies.

It is not easy to determine quantitative conclusions about these estimated matrices, since they contain so many values. So to assist us in such comparisons we have adopted the same several distinct scalar performance measures used in RP98. Accordingly, we have evaluated:

$$\left|\frac{F'F}{N}\right|, \quad \left[tr\left(\frac{F'F}{N}\right)\right]^{\frac{1}{2}}, \text{ and } \quad \left|\frac{F'F}{N} - I_m\right|,$$

for the true, PS89, and CBFA estimators. (Note that $[tr(\frac{F'F}{N})]^{\frac{1}{2}}$ denotes the norm of the $\frac{F'F}{N}$ matrix.) We have also differenced the matrix estimators pairwise, for the factor scores, factor loadings, and disturbance covariance matrices, to form:

$$\Delta F_{TP} = F_T - F_P, \quad \Delta F_{TC} = F_T - F_C,$$

$$\Delta\Lambda_{TP} = \Lambda_T - \Lambda_P, \quad \Delta\Lambda_{TC} = \Lambda_T - \Lambda_C, \quad \text{and}$$

$$\Delta\Psi_{TP} = \Psi_T - \Psi_P, \Delta\Psi_{TC} = \Psi_T - \Psi_C,$$

where T, P, and C denote the true values, the PS89 conditional modal, and CBFA estimators respectively.

Moreover, we have computed the scalar measures of the differenced matrices:

$$|(\Delta F')(\Delta F)|^{\frac{1}{2}}, \quad |(\Delta\Lambda')(\Delta\Lambda)|^{\frac{1}{2}}, \quad |(\Delta\Psi')(\Delta\Psi)|^{\frac{1}{2}},$$

and have compared their numerical values. All these comparisons are displayed in Tables 11 and 12. The $\frac{F'F}{N}$ matrices themselves are given below for the three types of estimators.

CBFA Matrix

$$\frac{\bar{F}'\bar{F}}{N} = \begin{pmatrix} 0.779 & 0.213 & 0.255 & 0.037 \\ 0.213 & 1.252 & 0.262 & -0.237 \\ 0.255 & 0.262 & 1.155 & 0.328 \\ 0.037 & -0.237 & 0.328 & 1.124 \end{pmatrix},$$

PS89 Matrix

$$\frac{\tilde{F}'\tilde{F}}{N} = \begin{pmatrix} 0.776 & 0.239 & 0.287 & 0.035 \\ 0.239 & 1.270 & 0.306 & -0.269 \\ 0.287 & 0.306 & 1.176 & 0.344 \\ 0.035 & -0.269 & 0.344 & 1.132 \end{pmatrix},$$

True Matrix

$$\frac{F_*'F_*}{N} = \begin{pmatrix} 1.079 & 0.048 & -0.095 & 0.037 \\ 0.048 & 1.455 & 0.211 & -0.328 \\ -0.095 & 0.211 & 1.357 & 0.253 \\ 0.037 & -0.328 & 0.253 & 0.981 \end{pmatrix}.$$

We note from inspection of Table 11 that the CBFA estimator of $\frac{F'F}{N}$ is closer to the ideal values for all of the three of the measures of performance which is better than the PS89 estimators. Since the matrix $\frac{F'F}{N}$ represents the sample covariance for the factor scores, and they have been generated from an identity covariance matrix the determinant should be the determinant of an identity covariance matrix, which is 1. That's what is meant by the ideal value in the last column of the table. Similarly for the norm of this matrix in the middle row since the square root of the

Table 11: Measures of Quality.

Performance Measures	CBFA Estimation	PS89 Estimation	True Values	Ideal Values
$\left\| \frac{F'F}{N} \right\|$	0.918	0.883	1.722	1
$\left[tr \left(\frac{F'F}{N} \right) \right]^{\frac{1}{2}}$	2.076	2.084	2.207	2
$\left\| \frac{F'F}{N} - I_m \right\|$	3.619×10^{-2}	3.637×10^{-2}	9.849×10^{-3}	0

trace of the identity matrix of order 4 is 2.

The difference measures in Table 12 clearly show that the factor scores, loadings, and disturbance variance and covariance estimates of CBFA are better than those of PS89. This is especially true for the loading and disturbance covariance matrices. We attribute this to CBFA yielding a better estimate of the sample

Table 12: Difference Measures Of Quality.

Performance Measures	CBFA-True	PS89-True	Ideal Values
$\|(\Delta F')(\Delta F)\|^{\frac{1}{2}}$	40.065	42.833	0
$\|(\Delta \Lambda')(\Delta \Lambda)\|^{\frac{1}{2}}$	52.695	63.036	0
$\|(\Delta \Psi')(\Delta \Psi)\|^{\frac{1}{2}}$	3.286×10^{9}	7.9385×10^{10}	0

covariance matrix $\hat{\Sigma}$ than PS89. This sample covariance matrix is the basis for the entire analysis.

7.2 Plankton Example

For our second example, we factor analyze the data given in Table 13 taken from Basilevsky, 1994 and originally from Imbrie and Kipp, 1971. The data consists of a core sample taken from the ocean floor.

Table 13: Plankton Data

N	1	2	3	4	5	6	7	8	9	10
1	1.792	0.489	43.485	0.814	25.570	0.651	0.000	0.163	0.000	0.163
2	3.203	0.712	37.722	0.356	30.961	0.712	0.000	0.356	0.000	0.000
3	2.364	1.709	47.009	0.855	20.513	1.709	0.000	1.282	0.427	0.000
4	1.124	0.562	47.191	1.124	12.360	2.247	0.000	3.933	0.562	0.562
5	0.671	1.007	43.624	3.020	15.436	1.007	0.000	0.336	0.671	0.336
6	1.149	0.766	52.874	0.766	12.261	0.000	0.000	0.383	2.299	0.000
7	1.990	0.498	53.234	3.980	6.965	0.000	0.000	0.498	0.995	0.000
8	2.222	2.222	45.926	2.222	13.333	2.963	0.000	1.481	1.481	1.481
9	1.786	1.190	49.405	1.786	10.714	1.786	0.000	0.595	0.595	0.000
10	0.621	0.621	36.025	2.484	10.519	0.621	0.000	1.242	1.863	0.000
11	1.418	0.000	46.099	2.837	9.220	4.255	0.000	0.709	2.836	0.000
12	0.000	0.000	38.298	0.709	11.348	2.837	0.000	1.418	5.674	0.000
13	0.498	0.498	48.756	0.000	5.970	1.990	0.498	0.498	2.985	0.000
14	1.379	1.034	42.069	0.690	8.621	2.069	0.000	2.759	1.724	0.690
15	0.662	0.000	46.358	0.000	11.921	0.000	0.000	1.987	3.311	0.000
16	3.429	1.143	45.714	1.143	14.286	1.714	0.000	0.571	3.429	0.571
17	2.899	2.899	42.995	0.000	14.010	1.449	0.000	2.415	2.415	0.483
18	1.198	1.796	50.299	1.198	8.383	2.994	0.000	0.599	0.599	0.599
19	1.887	2.516	38.994	3.145	7.547	2.516	0.000	1.258	1.258	0.000
20	5.143	2.857	38.286	0.000	13.714	1.143	0.000	1.143	1.143	0.000
21	3.067	0.613	37.423	1.227	13.497	2.761	0.000	1.227	0.000	0.307
22	1.961	2.614	41.830	3.268	11.765	1.307	0.654	1.307	0.654	0.000
23	1.515	2.020	37.374	1.010	12.626	2.020	0.000	0.000	0.505	0.000
34	1.422	2.844	38.389	1.422	16.114	0.948	0.000	0.000	0.474	0.000
25	1.630	1.630	36.957	2.174	10.870	2.174	0.000	0.000	0.000	0.000
26	1.571	1.571	37.696	1.571	10.995	4.188	0.000	2.094	2.618	1.047
27	1.826	3.196	36.073	0.913	12.329	2.283	0.000	0.457	0.913	0.457
28	0.926	3.241	28.241	0.463	12.037	0.926	0.000	0.463	1.852	0.463
29	1.379	2.414	35.517	0.345	11.679	0.345	0.000	0.000	4.828	0.000
30	1.036	6.218	34.197	1.036	14.508	0.518	0.000	0.000	1.554	0.518
31	0.649	3.896	39.610	3.896	13.636	1.299	0.000	0.543	0.649	0.000
32	1.485	7.426	29.208	2.475	15.842	1.485	0.000	2.970	1.485	0.000
33	1.087	0.000	42.391	1.630	15.761	1.630	0.000	2.174	1.087	0.000
34	3.404	0.426	32.766	4.255	13.191	2.128	0.000	3.830	0.851	1.700
35	1.429	0.476	42.381	2.857	10.952	1.905	0.000	0.476	0.952	1.900
36	1.449	3.623	36.957	0.000	15.942	3.623	0.000	0.725	1.449	0.720
37	1.685	1.685	48.315	2.809	10.674	1.124	0.000	1.124	1.124	0.000
38	0.772	0.386	40.927	0.772	15.444	2.703	0.000	0.000	0.772	0.380
39	1.266	1.266	37.975	2.532	18.143	3.376	0.000	2.110	0.422	0.000
40	3.627	0.518	41.451	1.554	16.580	0.518	0.000	2.591	1.554	0.000
41	1.869	1.402	37.850	2.804	12.617	2.336	0.000	9.813	0.467	0.930
42	3.509	2.456	42.105	2.105	12.281	1.053	0.351	2.456	0.000	0.000
43	0.904	0.904	44.578	1.205	14.759	0.602	0.301	1.506	0.602	0.000
44	1.449	0.483	43.961	3.865	12.560	1.449	0.000	2.899	0.000	0.000
45	1.299	0.649	38.961	0.325	17.208	1.945	0.000	4.545	1.948	0.000
46	0.000	0.741	33.333	2.222	22.222	2.222	0.000	0.741	0.000	0.000
47	2.513	4.523	35.176	1.005	20.603	0.000	0.000	0.000	0.000	0.000
48	1.026	0.513	42.051	2.051	16.410	2.051	0.000	0.513	2.051	0.000
49	0.565	0.565	44.068	3.955	10.169	1.695	0.000	9.605	3.390	0.000
50	1.523	0.000	34.518	2.030	20.305	2.030	0.000	1.523	1.015	0.000
51	0.508	0.000	40.609	0.508	21.827	0.508	0.000	3.046	0.000	0.000
52	0.000	2.703	28.649	1.622	24.324	3.784	0.000	2.162	3.243	0.000
53	0.629	4.403	39.623	0.629	10.063	3.145	0.000	5.660	5.031	0.000
54	0.800	2.400	50.400	1.600	11.200	2.400	0.000	4.800	0.000	0.000
55	1.630	0.543	54.348	2.174	7.609	3.804	0.000	1.630	2.717	0.000

In the core sample, plankton content is measured for $p = 10$ species at 110 depths. The plankton content is used to infer approximate climatological conditions which existed on earth. Since many species coexist at different times (core depths), their abundance is generally correlated, and a factor analysis is frequently performed. We take the first 55 observations as our training data and analyze the last $N = 55$

Table 12 Continued: Plankton Data

N	1	2	3	4	5	6	7	8	9	10
56	0.000	0.543	32.609	1.087	11.413	4.891	0.000	3.804	2.717	0.000
57	1.622	1.081	32.973	2.162	11.892	3.784	0.000	9.780	0.541	0.000
58	1.762	0.000	33.921	0.000	16.740	2.643	0.000	9.251	2.643	0.000
59	1.418	0.000	36.879	0.709	11.348	4.255	0.000	4.965	4.965	0.709
60	1.136	2.841	49.432	2.273	11.932	2.273	0.000	0.568	0.000	0.000
61	0.893	3.561	33.036	5.357	13.393	2.679	0.000	4.464	0.893	0.893
62	3.636	1.212	35.758	2.424	6.061	6.061	0.000	3.030	0.000	0.000
63	3.448	1.478	29.064	3.448	14.778	4.433	0.000	2.955	0.000	0.000
64	1.342	2.685	34.228	3.356	12.081	2.685	0.000	2.685	4.027	0.000
65	4.435	2.419	33.468	0.806	17.742	3.226	0.000	0.000	4.032	0.000
66	2.158	2.158	34.532	2.158	15.826	5.036	0.000	0.719	2.158	0.000
67	0.000	4.545	38.636	0.000	15.152	1.515	0.000	2.273	2.273	0.758
68	1.235	0.000	41.975	0.000	12.346	1.852	7.407	0.617	2.469	0.000
69	1.508	1.508	38.191	0.503	3.518	1.508	4.523	1.508	2.010	0.503
70	3.550	2.367	47.337	2.367	5.917	10.059	0.000	0.000	0.592	0.000
71	5.344	0.000	39.695	1.527	13.740	6.870	0.000	0.763	0.000	0.000
72	5.455	0.606	43.636	1.818	10.303	7.273	1.212	0.605	0.000	0.000
73	0.000	0.000	38.095	3.571	4.762	9.524	0.000	3.571	0.000	1.190
74	2.609	1.304	33.043	1.739	9.130	3.913	0.870	3.478	0.435	0.000
75	1.604	1.604	33.690	0.000	19.251	2.139	0.000	3.209	3.209	0.535
76	1.899	0.000	34.177	2.532	12.025	4.430	0.633	2.532	1.266	0.000
77	2.041	0.816	36.327	2.041	20.000	2.449	0.000	2.449	1.224	0.408
78	0.595	2.976	50.000	0.000	7.738	6.548	0.000	2.381	0.595	0.000
79	0.000	6.130	35.249	0.000	10.728	0.000	0.000	0.383	0.383	0.000
80	0.372	5.576	37.918	0.372	15.613	0.743	0.000	0.000	0.372	0.000
81	3.582	5.373	38.209	0.896	17.015	0.896	0.000	0.000	0.896	0.299
82	2.362	2.362	36.220	3.150	14.173	1.969	0.000	0.787	1.575	0.000
83	2.105	4.211	26.842	1.053	13.684	4.737	0.526	5.263	2.105	0.000
84	2.381	3.175	32.143	1.190	17.460	1.587	0.000	0.397	1.190	0.000
85	0.455	0.909	37.273	0.455	24.091	3.182	0.000	0.455	0.455	0.909
86	0.858	3.863	31.760	1.717	21.888	7.296	0.000	4.721	0.858	0.000
87	2.769	1.231	43.385	1.231	2.769	4.000	0.000	6.462	3.077	0.000
88	0.658	1.316	52.632	0.000	3.289	1.974	0.000	3.947	0.658	0.000
89	3.448	0.575	35.632	1.149	14.368	0.000	0.000	4.598	0.575	0.000
90	1.689	0.676	26.689	2.027	8.108	4.392	0.338	13.176	2.027	1.689
91	1.533	0.000	35.249	0.383	9.195	2.682	1.533	13.793	1.533	0.000
92	1.064	0.000	40.957	1.596	6.915	2.660	0.000	3.723	2.660	0.000
93	1.394	0.348	36.585	1.045	8.014	3.833	0.000	6.969	1.394	0.000
94	0.000	0.000	35.533	1.015	13.706	7.614	0.000	3.553	0.000	0.000
95	1.970	2.463	39.901	0.493	15.764	3.941	0.000	0.985	0.493	0.493
96	1.471	2.206	34.559	2.941	15.441	1.471	0.000	0.000	0.735	0.000
97	1.613	0.403	42.742	1.210	16.129	2.823	0.000	2.823	0.403	0.000
98	0.000	0.498	44.776	2.488	19.900	0.995	0.000	1.990	0.995	0.498
99	0.448	0.448	40.359	4.484	12.556	2.242	0.000	6.278	0.897	0.000
100	2.717	0.000	32.065	3.261	15.761	1.087	0.000	6.522	1.087	0.000
101	1.887	1.887	34.906	1.415	12.264	1.415	0.000	3.302	1.415	0.472
102	1.342	2.013	24.161	3.356	11.409	1.342	0.000	9.396	0.000	0.671
103	1.633	0.816	24.898	2.449	6.531	0.408	0.000	12.245	2.041	0.000
104	1.548	0.310	31.269	1.548	9.288	0.000	0.000	9.288	4.644	0.000
105	1.093	0.546	31.694	1.639	14.208	0.000	0.000	19.672	4.372	0.000
106	2.183	1.747	33.188	0.437	13.974	0.437	0.000	4.367	1.747	1.747
107	1.878	0.469	24.883	1.878	14.085	1.408	0.000	9.390	0.939	0.000
108	2.286	2.286	37.143	1.714	8.000	1.714	0.000	8.000	4.571	0.000
109	3.911	2.793	32.961	1.117	14.525	1.117	0.000	2.793	0.559	0.000
110	0.658	0.658	34.868	4.605	15.789	1.316	0.000	3.947	1.974	0.000

observations.

We have calculated the training sample mean $\bar{x}$ and the sample covariance matrix

Table 14: Plankton Training Mean.

p	mean
1	1.58
2	1.65
3	41.10
4	1.66
5	14.01
6	1.80
7	0.03
8	1.76
9	1.43
10	0.24

$\hat{\Sigma}$ which are given in Tables 14 and 15. From a principal components analysis we determined the number of factors m to be 4, 5, or 6 because they had eigenvalues 1.128, 1.064, and 0.842 respectively. These factors also accounted for 59.8%, 70.4%, and 78.8% of the variance respectively.

The prior $p(4) = p(5) = p(6) = \frac{1}{3}$ was assessed. Our prior means for the factor loadings hyperparameter is given in Table 17.

We assessed the remaining hyperparameters using method (a) as in the appendix. They are $h_0 = 55$, $\nu = 77$, and $b_0 = 174.5$.

We will be able to compare PS89 estimates to CBFA Gibbs sampling estimates.

Upon implementing the aforementioned, we found the number of factors to

Table 15: Plankton Sample Covariance matrix.

	1	2	3	4	5	6	7	8	9	10
1	1.01	0.12	0.10	-0.04	0.40	-0.18	0.00	-0.14	-0.35	0.08
2		2.43	-3.98	-0.17	0.06	-0.10	0.00	-0.20	0.01	-0.01
3			35.90	0.45	-11.59	-0.31	0.10	0.08	0.01	-0.23
4				1.34	-1.32	0.11	0.00	0.54	-0.37	0.06
5					23.03	-0.94	-0.11	-1.15	-2.09	-0.32
6						1.18	-0.02	0.38	0.29	0.13
7							0.02	-0.01	-0.01	-0.01
8								4.11	0.37	0.09
9									1.80	-0.03
10										0.20

be $\bar{m} = 5$ as evidenced in the following table containing the log of the conditional posterior probabilities (with an additive constant).

Number of Factors	$m = 4$	$m = 5$	$m = 6$
Log of Posterior	-7330.5	-5407.1	-6713.

Table 16: Simulation Posterior distribution for the number of factors.

Thus, CBFA correctly determined the number of factors to be five.

We found the estimate of the correlation parameter to be $\bar{\rho} = 0.5473$.

Table 17: Plankton Factor Loadings

Prior Factor Loadings, Λ_0.

p	1	2	3	4	5
1	0.068	-0.066	0.000	-0.518	-0.023
2	0.014	-1.210	-0.089	-0.029	0.141
3	3.193	3.150	0.022	-0.144	-0.107
4	0.137	0.059	0.185	-0.038	0.803
5	-3.375	-0.295	-0.672	-1.109	-0.789
6	-0.011	-0.034	0.135	0.549	0.275
7	0.037	-0.004	-0.010	-0.015	0.004
8	-0.062	0.112	1.050	0.237	0.444
9	0.355	-0.298	0.388	0.772	-0.360
10	0.023	-0.074	0.028	0.054	0.170

PS89 Factor Loadings, $\hat{\Lambda}$.

p	1	2	3	4	5
1	0.118	-0.191	0.000	-0.540	0.041
2	0.158	-1.002	-0.467	-0.309	-0.117
3	3.297	3.679	-0.270	-0.363	-0.366
4	-0.033	0.068	0.011	-0.139	0.934
5	-3.709	0.137	-0.921	-1.390	-1.094
6	0.135	0.320	-0.331	0.817	0.466
7	0.146	-0.008	-0.028	0.030	-0.060
8	-0.359	0.075	2.271	0.413	0.699
9	0.187	-0.227	0.431	0.747	-0.651
10	-0.013	-0.039	-0.016	0.035	0.131

CBFA Loadings, $\bar{\Lambda}$.

p	1	2	3	4	5
1	0.004	-0.103	-0.053	-0.364	-0.065
2	0.071	-0.695	-0.145	-0.100	0.024
3	3.150	3.569	-0.254	-0.319	-0.299
4	-0.029	-0.073	0.120	0.023	0.548
5	-3.372	0.152	-0.690	-1.114	-0.832
6	0.140	0.142	0.112	0.524	0.350
7	0.051	0.019	0.004	-0.008	-0.035
8	-0.346	-0.186	1.420	0.404	0.556
9	0.215	-0.175	0.295	0.492	-0.287
10	-0.012	-0.068	-0.011	0.045	0.062

Table 18: Plankton Factor Scores

PS89 Factor Scores, $\hat{F}$.

N	1	2	3	4	5
1	-0.194	-0.507	0.073	0.894	0.023
2	-0.457	-0.237	1.241	0.263	0.753
3	-1.101	0.375	1.092	0.077	-0.326
4	0.123	0.156	0.417	0.794	-0.405
5	1.274	1.829	-1.190	-0.933	-0.467
6	-0.299	-0.582	0.022	-0.105	0.810
7	0.971	-0.609	-0.092	0.424	0.805
8	-0.822	-0.906	-0.235	-0.105	0.577
9	0.106	-0.603	-0.144	0.358	-0.016
10	-0.614	-0.370	-0.851	-0.287	-0.964
11	-0.460	-0.134	-0.794	0.012	-0.281
12	0.093	0.152	-0.662	-0.294	-0.801
13	0.602	0.902	-0.847	-0.168	-0.930
14	1.780	-0.702	-0.358	0.508	0.010
15	1.914	1.045	-0.985	0.292	0.310
16	0.106	0.785	-0.844	-0.415	-0.134
17	0.963	1.011	-0.869	-0.362	0.009
18	0.998	0.022	-0.007	1.248	1.255
19	0.334	-0.791	-0.041	0.281	0.462
20	-1.098	0.111	-0.305	-0.209	-0.939
21	-0.122	-0.176	-0.270	0.280	0.220
22	-1.101	0.766	-0.609	-0.698	-0.556
23	1.792	1.631	-0.692	0.041	-0.329
24	0.770	-1.070	-1.028	-0.297	-0.373
25	0.152	-0.104	-1.289	-0.755	-0.726
26	0.031	0.002	-1.227	-1.169	-0.796
27	0.007	-0.088	-0.787	-0.415	-0.150
28	-0.718	-1.694	0.331	0.509	0.277
29	-0.689	-0.578	-0.965	-0.547	-0.542
30	-1.636	1.283	-1.294	-0.875	-1.117
31	-1.808	0.051	-0.224	-0.105	-0.098
32	1.959	0.298	0.752	0.637	0.291
33	2.633	1.800	-0.215	-0.153	-0.292
34	-0.260	0.140	0.028	-0.690	-0.240
35	-0.486	-1.469	2.322	1.186	1.358
36	-0.061	0.083	2.208	0.511	0.621
37	1.161	0.401	0.064	0.457	-0.035
38	0.513	-0.072	0.731	0.553	0.389
39	-0.494	0.388	-0.269	0.545	0.181
40	-0.037	0.698	-0.989	-0.528	-0.595
41	-0.303	-0.206	-1.031	-0.484	-0.185
42	-0.109	1.531	-0.601	-0.694	-0.566
43	-0.519	2.180	-0.919	-1.065	-0.868
44	0.065	0.988	0.361	-0.215	0.471
45	-0.939	-0.089	0.541	-0.355	0.289
46	0.088	-0.345	-0.188	-0.129	-0.103
47	-0.879	-1.852	1.305	0.359	1.274
48	-0.195	-2.037	2.214	0.933	1.279
49	-0.015	-0.894	1.521	0.718	0.173
50	-1.262	0.032	3.576	0.443	0.568
51	-0.307	-0.459	0.039	-0.221	-0.296
52	-1.303	-1.340	1.301	0.293	0.668
53	0.761	-0.361	1.113	0.502	0.090
54	-0.256	-0.624	-0.371	-0.627	-0.193
55	-0.656	0.239	-0.061	-0.215	0.163

CBFA Factor Scores, $\bar{F}$.

N	1	2	3	4	5
1	-0.175	-0.384	-0.011	0.688	-0.093
2	-0.307	-0.086	0.627	0.079	0.432
3	-0.847	0.366	0.593	0.052	-0.351
4	0.127	0.146	0.294	0.661	-0.375
5	0.996	1.396	-0.454	-0.589	-0.036
6	-0.230	-0.511	-0.012	-0.116	0.645
7	0.769	-0.445	-0.086	0.218	0.552
8	-0.666	-0.720	-0.233	-0.135	0.390
9	0.126	-0.536	-0.051	0.335	-0.014
10	-0.354	-0.393	-0.422	-0.058	-0.651
11	-0.515	-0.225	-0.422	0.105	-0.134
12	0.073	0.010	-0.352	-0.157	-0.450
13	0.425	0.739	-0.458	-0.048	-0.602
14	1.479	-0.531	-0.203	0.328	-0.056
15	1.477	0.779	-0.332	0.265	0.423
16	0.011	0.613	-0.406	-0.292	0.007
17	0.729	0.794	-0.357	-0.260	0.160
18	0.792	0.095	0.004	0.871	0.890
19	0.263	-0.595	-0.112	0.133	0.255
20	-0.944	0.008	-0.175	-0.045	-0.680
21	-0.126	-0.116	-0.195	0.200	0.134
22	-0.931	0.557	-0.310	-0.438	-0.294
23	1.418	1.255	-0.160	0.124	-0.044
24	0.570	-0.989	-0.636	-0.189	-0.231
25	0.057	-0.254	-0.699	-0.462	-0.379
26	-0.046	-0.181	-0.630	-0.783	-0.396
27	-0.029	-0.138	-0.425	-0.262	-0.010
28	-0.600	-1.431	0.063	0.344	0.059
29	-0.633	-0.580	-0.601	-0.350	-0.346
30	-1.444	0.903	-0.688	-0.480	-0.640
31	-1.549	-0.116	-0.122	0.009	-0.010
32	1.673	0.337	0.537	0.425	0.168
33	2.184	1.521	0.079	-0.095	-0.065
34	-0.198	0.161	-0.032	-0.559	-0.181
35	-0.271	-1.005	1.133	0.683	0.676
36	0.090	0.259	1.187	0.225	0.237
37	0.973	0.395	0.097	0.345	-0.050
38	0.473	0.043	0.388	0.341	0.164
39	-0.460	0.318	-0.176	0.422	0.105
40	-0.106	0.465	-0.489	-0.306	-0.295
41	-0.293	-0.236	-0.609	-0.314	-0.047
42	-0.128	1.217	-0.236	-0.447	-0.261
43	-0.464	1.695	-0.353	-0.637	-0.367
44	0.100	0.863	0.273	-0.171	0.433
45	-0.724	0.002	0.228	-0.327	0.166
46	0.073	-0.282	-0.139	-0.103	-0.088
47	-0.638	-1.382	0.491	0.070	0.702
48	-0.007	-1.429	1.003	0.451	0.571
49	0.112	-0.592	0.776	0.464	-0.103
50	-0.791	0.234	1.969	0.193	0.147
51	-0.239	-0.367	-0.045	-0.170	-0.262
52	-0.995	-0.960	0.509	0.066	0.244
53	0.720	-0.242	0.682	0.345	-0.025
54	-0.230	-0.544	-0.271	-0.499	-0.141
55	-0.519	0.200	-0.036	-0.133	0.176

Table 19: Plankton Disturbance Covariance Matrices

PS89 Disturbance Covariance Matrix, $\hat{\Psi}_{mode}$.

	1	2	3	4	5	6	7	8	9	10
1	1.88	-0.20	-0.12	-0.01	-0.12	0.31	-0.00	-0.13	-0.01	-0.05
2		2.03	0.42	-0.03	0.35	0.01	-0.14	-0.20	0.05	0.03
3			1.70	0.01	0.38	0.18	-0.06	-0.15	0.11	0.03
4				1.73	0.10	0.03	-0.10	-0.26	0.04	-0.01
5					1.78	0.10	-0.09	-0.07	0.22	0.05
6						3.11	-0.10	-0.64	-0.29	-0.01
7							1.81	-0.06	0.02	0.00
8								2.81	0.05	-0.03
9									1.76	0.02
10										1.35

CBFA Disturbance Covariance Matrix, $\bar{\Psi}$.

	1	2	3	4	5	6	7	8	9	10
1	3.05	-0.17	-0.23	-0.30	0.11	-0.03	0.08	-0.36	-0.14	-0.19
2		3.13	0.68	-0.22	0.58	-0.13	-0.44	-0.53	-0.12	0.09
3			9.90	-0.56	-0.06	0.51	-0.21	-2.48	0.06	-0.00
4				2.79	-0.30	0.22	-0.17	0.17	-0.16	0.01
5					8.51	-0.76	-0.23	-1.31	-0.07	0.06
6						5.26	-0.37	0.01	-0.44	0.03
7							2.51	0.09	0.05	-0.10
8								8.88	0.42	-0.20
9									3.04	0.07
10										1.63

Using CBFA, we identify there to be five underlying factors. The factors (in the order 1, 2, 5, 3, 4) correspond to the four climate zones Tropical, Subtropical, Polar, Subpolar, and the Gyre margin assemblege region.

We see that factor two is more defined to load on variable two in the PS89 estimate and more so in the CBFA estimate of the loading matrix and factor four is more defined to load on variable five.

Notice how the CBFA estimates of the disturbance variances are all larger than the PS89 and some are several times larger. This is because the PS89 estimates are inefficient when the observations are correlated.

8 Conclusion and Future Research

We have developed the full covariance correlated Bayesian factor analysis model; we have shown that the number of parameters is enormous; we have shown that the number of covariance parameters can be greatly reduced by specifying either a separable or a matrix intraclass covariance matrix; and we have shown that that the separable covariance matrix can be estimated but the computational requirements are too great. We have simplified the computation by considering a homoscedastic variance model decomposing Ω into Φ and Ψ where Φ is a correlation matrix that is a function of the single parameter ρ and Ψ is a covariance matrix. We have outlined a couple of correlation structures for Φ and have used a first order Markov correlation matrix in two examples.

We see that the addition of a single parameter models correlation vectors and increases how well we explain the relationship among a set of observed random variables in terms of a smaller number of factors as measured several scalar performance measures.

It is our conclusion that when correlated observations are factor analyzed as independent observations, the covariance matrix Σ is inefficiently estimated and thus the factor analysis based on this covariance matrix can be improved.

Future research will include investigation of the independent vector assumption when in fact the observation vectors are correlated.

We csn investigate other covariance structures and covariances that depend on more than one parameter. We have laid the foundation for the correlation matrix

Φ to depend on several parameters, but for simplicity consider it to be a function of a single parameter.

We have evaluated examples that assumed orthogonal factors. We can consider examples with oblique factors.

We might use sampling distributions other than the normal such as the multivariate t distribution.

Finally, future research will include applications to imaging science, economics, and to the social sciences. In imaging, we can factor analyze an image taken in several different modalities or bands into a smaller number of “common” factor images that represent the essence of the images. These “common” factor images can be stored thus a lossy compression technique or can be used for object recognition. In Economics, we can use factor analysis for portfolio selection. In the social sciences, we can use correlated Bayesian factor analysis when it is found that the subjects (observation vectors) are related.

Appendices

A Bayesian Estimation Methods

In this section we define define the estimation procedures used and discuss some advantages of each. The procedures are marginalization and conditional estimation, LSO, and Gibbs sampling

A.1 Conditional Modal Estimation

Often we have a set of parameters, $\theta = (\theta_1, \ldots, \theta_J)$ in our posterior distribution $p(\theta|X)$. The marginal posterior distribution of any of the parameters, say θ_j can be obtained by integrating $p(\theta|X)$ with respect all parameters except θ_j. That is

$$p(\theta_j|X) = \int p(\theta_1, \ldots, \theta_J)\ d\theta_1 \ldots d\theta_{j-1}\ d\theta_{j+1} \ldots d\theta_J \tag{A.1.1}$$

where the integral is evaluated over the appropriate range of the parameters. It is possible to calculate the marginal posterior distribution for each of the parameters and calculate marginal posterior estimates such as the mean

$$\hat{\theta}_j = E(\theta_j|X) = \int \theta_j p(\theta_j|X) d\theta_j. \tag{A.1.2}$$

We may instead choose to compute conditional posterior distributions. If again $\theta = (\theta_1, \ldots, \theta_J)$, then the conditional distribution of any one of the parameters say θ_k given another say θ_j is given by

$$p(\theta_k|\theta_j, X) = \frac{p(\theta_k, \theta_j|X)}{p(\theta_j|X)} \tag{A.1.3}$$

where

$$p(\theta_k, \theta_j|X) = \int p(\theta_1, \ldots, \theta_J|X) d\theta_1 \ldots d\theta_{j-1}\ d\theta_{j+1} \ldots d\theta_{k-1}\ d\theta_{k+1} \ldots d\theta_J. \tag{A.1.4}$$

We may now compute conditional posterior mean (and mode) estimators such as

$$\hat{\theta}_k = E(\theta_k|\theta_j, X) = \int \theta_k p(\theta_k|\theta_j, X) d\theta_k. \tag{A.1.5}$$

or

$$\hat{\theta}_l = E(\theta_l|\theta_k, \theta_j, X) = \int \theta_l p(\theta_l|\theta_k, \theta_j, X) d\theta_l. \tag{A.1.6}$$

where

$$p(\theta_l|\theta_k, \theta_j, X) = \frac{p(\theta_l, \theta_k, \theta_j|X)}{p(\theta_k, \theta_j|X)}. \tag{A.1.7}$$

A.2 Lindley/Smith Optimization (LSO)

Lindley/Smith optimization (Lindley and Smith, 1972) is a deterministic optimization method that finds the joint posterior modal estimators of $p(\theta|X)$ where θ denotes the vector of parameters, and X denotes the data.

Assume that $\theta = (\theta_1, \theta_2)$ where θ_1 and θ_2 are scalars and the posterior density of θ is $p(\theta_1, \theta_2|X)$. We have a surface in 3-Dimensional space. We have θ_1 along one axis and θ_2 along the other with $p(\theta_1, \theta_2|X)$ being the height of the surface or hill.

We want to find the top of the hill which is the same as finding the peak or maximum of the function $p(\theta_1, \theta_2|X)$ with respect to both θ_1 and θ_2. Well we find the maximum of a surface by differentiating with respect to each variable (direction).

The maximum of the function $p(\theta_1, \theta_2|X)$ satisfies

$$\frac{\partial}{\partial\theta_1} p(\theta_1, \theta_2|X) \Big|_{\theta_1=\tilde{\theta}_1} = \frac{\partial}{\partial\theta_2} p(\theta_1, \theta_2) \Big|_{\theta_2=\tilde{\theta}_2} = 0, \qquad \text{(A.2.1)}$$

which is the same as

$$\frac{\partial}{\partial\theta_1} p(\theta_1|\theta_2, X) p(\theta_2|X) \Big|_{\theta_1=\tilde{\theta}_1} = \frac{\partial}{\partial\theta_2} p(\theta_2|\theta_1, X) p(\theta_1|X) \Big|_{\theta_2=\tilde{\theta}_2} = 0 \qquad \text{(A.2.2)}$$

or

$$p(\theta_2|X) \frac{\partial}{\partial\theta_1} p(\theta_1|\theta_2, X) \Big|_{\theta_1=\tilde{\theta}_1} = p(\theta_1|X) \frac{\partial}{\partial\theta_2} p(\theta_2|\theta_1, X) \Big|_{\theta_2=\tilde{\theta}_2} = 0 \qquad \text{(A.2.3)}$$

assuming that $p(\theta_1|X) \neq 0$ and $p(\theta_2|X) \neq 0$.

We can obtain the posterior conditionals (functions) $p(\theta_1|\theta_2, X)$ and $p(\theta_2|\theta_1, X)$ along with their respective modes (maximum) $\tilde{\theta}_1 = \tilde{\theta}_1(\theta_2, X)$ and $\tilde{\theta}_2 = \tilde{\theta}_2(\theta_1, X)$.

We have the maximum of θ_1, $\tilde{\theta}_1$ for a given value of (conditional on) θ_2, and the maximum of θ_2, $\tilde{\theta}_2$ for a given value of (conditional on) θ_1.

The optimization procedure consists of
(1) Selecting an initial value for θ_2; call it $\tilde{\theta}_2^{(0)}$.
(2) Calculate the modal (maximal) value of $p(\theta_1|\tilde{\theta}_2^{(0)}, X)$, $\tilde{\theta}_1^{(1)}$.
(3) Calculate the modal (maximal) value of $p(\theta_2|\tilde{\theta}_1^{(1)}, X)$, $\tilde{\theta}_2^{(1)}$.
(4) Continue to calculate the remainder of the sequence $\tilde{\theta}_1^{(1)}, \tilde{\theta}_2^{(1)}, \tilde{\theta}_1^{(2)}, \tilde{\theta}_2^{(2)}, \ldots$ until convergence is reached.

If the posterior conditional distributions are not unimodal, we may converge to a local maximum and not the global maximum. If the posterior conditionals are unimodal, then we will always converge to a global maximum.

When convergence is reached, the point estimators $(\tilde{\theta}_1, \tilde{\theta}_2)$ are the maximum a posteori estimators.

This method can be generalized to more than two parameters. According to O'Hagan (1994),

> If θ is partitioned by $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$ into J groups of parameters, we begin with a starting point $\tilde{\theta}^{(0)} = (\tilde{\theta}_1^{(0)}, \tilde{\theta}_2^{(0)}, \ldots, \tilde{\theta}_J^{(0)})$ and at the i^{th} iteration define $\tilde{\theta}^{(i+1)}$ by
>
> $$\begin{aligned} \tilde{\theta}_1^{(i+1)} &= \tilde{\theta}_1(\tilde{\theta}_2^{(i)}, \tilde{\theta}_3^{(i)}, \ldots, \tilde{\theta}_J^{(i)}) && \text{(A.2.4)} \\ \tilde{\theta}_2^{(i+1)} &= \tilde{\theta}_2(\tilde{\theta}_1^{(i+1)}, \tilde{\theta}_3^{(i)}, \ldots, \tilde{\theta}_J^{(i)}) && \text{(A.2.5)} \\ &\vdots \\ \tilde{\theta}_J^{(i+1)} &= \tilde{\theta}_1(\tilde{\theta}_2^{(i+1)}, \tilde{\theta}_3^{(i+1)}, \ldots, \tilde{\theta}_{J-1}^{(i+1)}) && \text{(A.2.6)} \end{aligned}$$
>
> at each step computing the maximum or mode. To apply this method we need to determine the functions $\tilde{\theta}_j$ which give the maximum of $p(\theta|X)$ with respect to $\tilde{\theta}_j$, conditional on the fixed values of all the other elements of θ.

This is the general form of LSO.

A.3 Gibbs Sampling

Gibbs sampling is a stochastic method that draws random samples from the posterior conditional distribution for each of the parameters conditional on the fixed values of all the other parameters and the data X. Let $p(\theta|X)$ be the posterior distribution of the parameters where θ is the set of parameters and X is the data. Let θ is partitioned by $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$ into J groups of parameters. Ideally, we would like to perform the integration of the joint posterior distribution to obtain marginal posterior distributions

$$p(\theta_j|X) = \int p(\theta_1, \ldots, \theta_J)\ d\theta_1 \ldots d\theta_{j-1}\ d\theta_{j+1} \ldots d\theta_J \tag{A.3.1}$$

and marginal posterior mean estimates

$$E(\theta_j|X) = \int \theta_j p(\theta_j|X) d\theta_j. \tag{A.3.2}$$

Unfortunately, these integrations are usually of very high dimension and not available in a closed form. This is why we need the Gibbs sampling procedure. With the random samples drawn from the posterior conditional distributions, we can estimate the marginal posterior distributions and the marginal posterior means.

For the Gibbs sampling, we begin with an initial value

$$\bar{\theta}^{(0)} = (\bar{\theta}_1^{(0)}, \bar{\theta}_2^{(0)}, \ldots, \bar{\theta}_J^{(0)})$$

and at the i^{th} iteration define

$$\bar{\theta}^{(i+1)} = (\bar{\theta}_1^{(i+1)}, \bar{\theta}_2^{(i+1)}, \ldots, \bar{\theta}_J^{(i+1)})$$

by the values from

$$\begin{aligned}
\bar{\theta}_1^{(i+1)} &= \text{a random sample from } p(\bar{\theta}_1 | \bar{\theta}_2^{(i)}, \bar{\theta}_3^{(i)}, \ldots, \bar{\theta}_J^{(i)}, X) && \text{(A.3.3)} \\
\bar{\theta}_2^{(i+1)} &= \text{a random sample from } p(\bar{\theta}_2 | \bar{\theta}_1^{(i+1)}, \bar{\theta}_3^{(i)}, \ldots, \bar{\theta}_J^{(i)}, X) && \text{(A.3.4)} \\
&\vdots \\
\bar{\theta}_J^{(i+1)} &= \text{a random sample from } p(\bar{\theta}_J | \bar{\theta}_1^{(i+1)}, \bar{\theta}_2^{(i+1)}, \ldots, \bar{\theta}_{J-1}^{(i+1)}, X) && \text{(A.3.5)}
\end{aligned}$$

that is, at each step drawing a random sample from the conditional posterior distribution. To apply this method we need to determine the posterior conditionals of θ_j, conditional on the fixed values of all the other elements of θ and X from $p(\theta|X)$.

We will have $\bar{\theta}^{(1)}, \bar{\theta}^{(2)}, \ldots, \bar{\theta}^{(s+1)}, \ldots, \bar{\theta}^{(s+t)}$. The first s random samples called the "burn in" are discarded and the remaining t samples are kept.

The marginal posterior distributions (Equation A.3.1) are estimated by

$$\bar{p}(\theta_j) = \frac{1}{t} \sum_{k=1}^{t} p(\bar{\theta}_j^{(s+k)} | \bar{\theta}_1^{(s+k)}, \bar{\theta}_2^{(s+k)}, \cdots, \bar{\theta}_{j-1}^{(s+k)}, \bar{\theta}_{j+1}^{(s+k)}, \cdots, \bar{\theta}_J^{(s+k)}, X), \ j = 1, \ldots, J \qquad \text{(A.3.6)}$$

and the marginal posterior mean estimators of the parameters (Equation A.3.2) are estimated by $\bar{\theta} = (\bar{\theta}_1, \ldots, \bar{\theta}_J)$ where

$$\bar{\theta}_j = \frac{1}{t} \sum_{k=1}^{t} \bar{\theta}_j^{(s+k)}, \quad j = 1, \ldots, J. \tag{A.3.7}$$

A.4 Advantages of Marginalization And Conditional Estimation over LSO and Gibbs Sampling

The advantage of marginalization and conditional estimation over LSO and Gibbs sampling is that there are analytic equations for the estimators. This result eliminates the the need for a computationally intensive method thus dramatically reducing the computation time and convergence issues. However, this method could not be used because we cannot obtain marginal distributions in closed form when Φ has unknown parameters.

A.5 Advantages of LSO Over Gibbs Sampling

We will show that when Φ is known, each of the posterior conditional distributions are unimodal. Thus we do not have to worry about local maxima, we will find global maxima. The reason one would use a stochastic procedure like Gibbs sampling over a deterministic procedure like LSO is to eliminate the possibility of converging to a local mode when the conditional posterior distribution is multimodal.

LSO is slightly simpler to implement than Gibbs and less computationally intensive because Gibbs sampling requires generation of random samples from the conditionals. LSO simply has to cycle through the posterior conditional modes and convergence is not uncertain as it is with Gibbs sampling. With LSO, we can check for convergence say every 1000 iterations by computing the difference between $\theta_j^{(1000k)}$ and $\theta_j^{(1000(k+1))}$ for every j and if each element is the same to the 3^{rd} decimal, we can claim convergence and stop. This reduces computation time.

We do not implement LSO because when Φ has unknown parameters, the posterior conditional is not unimodal. LSO might converge to a local maxima.

A.6 Advantages of Gibbs Sampling Over LSO

When the posterior conditionals are not recognizable as unimodal distributions, we would want to use a stochastic procedure like Gibbs sampling to eliminate the possibility of converging to a local maxima. Although Gibbs sampling is more computationally intensive than LSO, it is a more general method and gives us more information such as marginal posterior estimates.

A.7 Gibbs Sampling Convergence

The Gibbs sampling procedure in the current form was introduced by Geman and Geman, 1984. Hastings, 1970 developed essentially the same idea.

It is well known that the full posterior conditional distributions uniquely determine the full joint density when the random variables have a joint distribution whose density function is strictly positive over the sample space (Gelfand and Smith, 1990). Since the posterior conditionals uniquely determine the full joint density, they also uniquely determine the posterior marginals. Geman and Geman showed that under mild conditions, the following results are true.

Result 1 *(Convergence)*

$$(\bar{\theta}_1^{(i)}, \bar{\theta}_2^{(i)}, \ldots, \bar{\theta}_J^{(i)}) \xrightarrow{d} (\theta_1, \theta_2, \ldots, \theta_J)$$

and hence for each j, $\bar{\theta}_j^{(i)} \xrightarrow{d} \theta_j \sim p(\theta_j)$ *as* $i \to \infty$.

Result 2 *(Rate)*
Using the sup norm, rather than the L_1 *norm, the joint density of* $(\bar{\theta}_1^{(i)}, \bar{\theta}_2^{(i)}, \ldots, \bar{\theta}_J^{(i)})$ *converges to the true joint density* $p(\theta_1, \theta_2, \ldots, \theta_J)$ *at a geometric rate in* i*, under visiting in the natural order. A minor adjustment to the rate is required for an arbitrary io visiting scheme.*

Result 3 *(Ergodic Theorem)*
For any measurable function T *of* $(\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_J)$ *whose expectation exists,*

$$s \overset{lim}{\rightarrow} \infty \sum_{s=1}^{t} T(\bar{\theta}_1^{(i)}, \bar{\theta}_2^{(i)}, \ldots, \bar{\theta}_J^{(i)}) \xrightarrow{a.s.} E(T(\theta_1, \theta_2, \ldots, \theta_J)).$$

With these results, we are guaranteed convergence.

B Covariance Determination

Here we describe a method to determine which correlation structure to determine. The possible structures are separable independent, first order Markov, intraclass and matrix intraclass.

B.1 Separable

$$\begin{array}{rl} \text{(I)} & (\epsilon|\Phi,\Psi) \sim N(0, \Phi\otimes\Psi) \\ \text{(II)} & (f|\Phi,m) \sim N(0,\Phi\otimes R) \\ \text{(III)} & (f|\Phi,m) \text{ and } (\epsilon|\Phi\otimes\Psi) \text{ are independent} \end{array} \tag{B.1.1}$$

From (I)–(III) above,

$$p(X|\Phi,\Psi,m,F,\Lambda) = (2\pi)^{-\frac{Np}{2}}|\Phi|^{-\frac{p}{2}}|\Psi|^{-\frac{N}{2}}e^{-\frac{1}{2}tr\Psi^{-1}(X-F\Lambda')'\Phi^{-1}(X-F\Lambda')} \tag{B.1.2}$$

and

$$p(F|\Phi,m) = (2\pi)^{-\frac{Nm}{2}}|R|^{-\frac{N}{2}}|\Phi|^{-\frac{m}{2}}e^{-\frac{1}{2}tr\Phi^{-1}FR^{-1}F'}. \tag{B.1.3}$$

Multiplying the above expressions we obtain $p(F,X|\Phi,m,\Lambda,\Psi)$ which can be integrated with respect to F to obtain

$$p(X|\Phi,m,\Lambda,\Psi) = (2\pi)^{-\frac{Np}{2}}|\Phi|^{-\frac{p}{2}}|\Lambda R\Lambda'+\Psi|^{\frac{N}{2}}e^{-\frac{1}{2}tr\Phi^{-1}X(\Lambda R\Lambda'+\Psi)^{-1}X'} \tag{B.1.4}$$

But we are not interested in Λ and Ψ so let $\Sigma = \Lambda R\Lambda' + \Psi$ and

$$p(X|\Phi, \Sigma) = (2\pi)^{-\frac{Np}{2}} |\Phi|^{-\frac{p}{2}} |\Sigma|^{-\frac{N}{2}} e^{-\frac{1}{2} tr \Phi^{-1} X \Sigma^{-1} X'}. \tag{B.1.5}$$

The maximum likelihood estimates are

$$\hat{\Sigma} = \frac{X' \hat{\Phi}^{-1} X}{N} \tag{B.1.6}$$

and

$$\hat{\Phi} = \frac{X \hat{\Sigma}^{-1} X'}{p}. \tag{B.1.7}$$

We cycle between the equations. But since $p < N$, $\hat{\Phi}$ is singular. We must consider structures for Φ.

B.2 Separable Markov

Cycle between

$$\hat{\Sigma} = \frac{X' \hat{\Phi}^{-1} X}{N} \tag{B.2.1}$$

and $\hat{\rho}$, where $\hat{\rho}$ is the max of

$$ln[p(X|\rho, \Sigma)] = -\frac{p(N-1)}{2} ln(1-\rho^2) - \frac{k_1}{2} \frac{1}{1-\rho^2} + \frac{k_2}{2} \frac{\rho}{1-\rho^2} - \frac{k_3}{2} \frac{\rho^3}{1-\rho^2} \tag{B.2.2}$$

and k_1, k_2, and k_3 are scalars that depend on $\hat{\Sigma}$.

$$\hat{\Phi} = \begin{pmatrix} 1 & \hat{\rho} & \hat{\rho}^2 & \cdots & \hat{\rho}^{N-1} \\ & 1 & \hat{\rho} & \cdots & \hat{\rho}^{N-2} \\ & & \ddots & & \vdots \\ & & & & 1 \end{pmatrix} \tag{B.2.3}$$

B.3 Separable Intraclass

Cycle between

$$\hat{\Sigma} = \frac{X'\hat{\Phi}^{-1}X}{N} \tag{B.3.1}$$

and $\hat{\rho}$, where $\hat{\rho}$ is the max of

$$\begin{aligned} ln[p(X|\rho,\Sigma)] &= -\frac{p(N-1)}{2}ln(1-\rho) - \frac{p}{2}ln[1+\rho(N-1)] \\ &- \frac{c_1}{2}\frac{1}{1-\rho} + \frac{c_2}{2}\frac{\rho}{(1-\rho)[1+\rho(N-1)]} \end{aligned} \tag{B.3.2}$$

and c_1 and c_2 are scalars that depend on $\hat{\Sigma}$.

$$\hat{\Phi} = \begin{pmatrix} 1 & \hat{\rho} & \hat{\rho} & \cdots & \hat{\rho} \\ & 1 & \hat{\rho} & \cdots & \hat{\rho} \\ & & \ddots & & \vdots \\ & & & & \hat{\rho} \\ & & & & 1 \end{pmatrix} \tag{B.3.3}$$

B.4 Separable Independent

$$\hat{\Sigma} = \frac{X'X}{N} \tag{B.4.1}$$

and

$$\hat{\Phi} = I_N. \tag{B.4.2}$$

B.5 Matrix Intraclass

$$\begin{array}{rl} \text{(I)} & (\epsilon|\Omega) \sim N(0,\Omega) \\ \text{(II)} & (f|\Theta,m) \sim N(0,\Theta) \\ \text{(III)} & (f|\Theta,m) \text{ and } (\epsilon|\Omega) \text{ are independent} \end{array} \tag{B.5.1}$$

where

$$\Omega = \begin{pmatrix} \Psi & \Upsilon & & \cdots & \Upsilon \\ & \Psi & & & \\ & & \ddots & & \vdots \\ & & & & \Upsilon \\ & & & & \Psi \end{pmatrix} \tag{B.5.2}$$

and

$$\Theta = \begin{pmatrix} R & P & & \cdots & P \\ & R & & & \\ & & \ddots & & \vdots \\ & & & & P \\ & & & & R \end{pmatrix}.$$

$$p(x|\Psi,\Upsilon,m,f,\Lambda) = (2\pi)^{-\frac{Np}{2}}|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}[x-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-(I_N\otimes\Lambda)f]}. \tag{B.5.3}$$

and

$$p(f|m,R,P) = (2\pi)^{-\frac{Nm}{2}}|\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}f'\Theta^{-1}f}. \tag{B.5.4}$$

Going through the orthogonal transformation, integrating with respect to f, and neglecting the first transformed observation we find

$$p(Z|m,\Lambda,\Xi,R_2) = (2\pi)^{-\frac{(N-1)p}{2}}|\Lambda R_2\Lambda' + \Xi|^{\frac{(N-1)}{2}} e^{-\frac{1}{2}tr\Phi^{-1}Z(\Lambda R_2\Lambda'+\Xi)^{-1}Z'} \tag{B.5.5}$$

But we are not interested in Λ and Ξ so let $\Sigma = \Lambda R_2\Lambda' + \Xi$ and

$$p(Z|\Phi,\Sigma) = (2\pi)^{-\frac{Np}{2}}|\Sigma|^{-\frac{(N-1)}{2}} e^{-\frac{1}{2}trX\Sigma^{-1}X'}. \tag{B.5.6}$$

The maximum likelihood estimate is

$$\hat{\Sigma} = \frac{Z'Z}{N-1}. \tag{B.5.7}$$

and

$$\hat{\Phi} = I_{N-1}. \tag{B.5.8}$$

B.6 Determination

Let the various covariance models be denoted by

M_1 = Separable Independent

M_2 = Separable Markov

M_3 = Separable Intraclass

$M_4 =$ Matrix Intraclass

We assess $p(M_i)$, i=1,...,4 then select M_i that makes

$$p(M_i|\Phi_i, \Sigma_i, Data) = p(M_i)p(Data|\Phi_i, \Sigma_i, M_i) \tag{B.6.1}$$

a maximum given $\Phi_i = \hat{\Phi}_i$ and $\Sigma_i = \hat{\Sigma}_i$. Note that $Data$ is either X or Z.

C Hyperparameter Assessment

In this section, we describe the process of assessing the hyperparameters of the prior distributions for the parameters under the separable model. Our methods are very simple and very easy to implement. For details and more elaborate methods see Hayashi, 1997 but we do not believe that more elaborate methods are necessary as demonstrated by Lee 1994, and Lee and Press, 1998.

We can either a) assume that there is previous data available or b) that there is not previous data available. We will discuss both. The previous data could be part of the current data set used for hyperparameter assessment. Let the previous data be Y with n observations

Hyperparameters for $p(m)$

a) We perform a principal components analysis on the sample covariance matrix $\hat{\Sigma}$ of the training data Y and to determine the range of values for the number of factors m. We do this by using Kaiser's eigenvalue-one criterion (Kaiser, 1960). In the Kaiser criterion you select the number of factors to be that number which has an eigenvalue greater than one. In this criterion, a factor is retained provided it explains at least as much of the variability as one test score. We select this number m_{E1} to be our most likely value and retain one more m_u and one less m_l. We assign the prior probability for the number of factors to be $p(m_l) = p(m_{E1}) = p(m_u) = \frac{1}{3}$.

b) The possible number of factors can be assigned purely subjectively and so can their distribution.

Hyperparameters for $p(\Lambda|\Psi, m)$

a) We assign Λ_0 by performing a classical, independent data vectors factor analysis on the training data Y and using the resulting factor loadings. We choose to perform a principal factor analysis.

We simplify the assessment of H by assuming that $H = h_0 I_m$. We assess the hyperparameter h_0 using the following method attributed to Hayashi. Maximum likelihood estimators are obtained by replacing $Var(\lambda)$ and $E(\Psi)$ by $Var(\hat{\lambda})$ and $\hat{\Psi}$ in

$$Var(\lambda) = E(\Psi) \otimes H^{-1}.$$

This assumes Normality.We also assume that

$$\sum_{i=1}^{p} Var(\hat{\lambda}_{ij}) = \sum_{i=1}^{p} Var(\hat{\lambda}_{ik}), \quad j \neq k$$

and

$$\prod_{i=1}^{p} Var(\hat{\lambda}_{ij}) = \prod_{i=1}^{p} Var(\hat{\lambda}_{ik}), \quad j \neq k$$

along with the large sample approximation

$$\frac{F'F}{N} \approx I_m$$

so that we assess the hyperparameter h_0 as

$$h_0 = n \ ,$$

where n is the training data sample size.

b) We may also use pure subjective prior experience where h_0 determines how much less variable the loadings are than an individual observation.

Hyperparameters for $p(\Psi)$

We simplify the assessment by assuming that hyperparameter B is $B = b_0 I_p$ where b_0 is a scalar constant.

a) We assess the hyperparameter ν, the prior degrees of freedom by a method due to Hayashi. We start with the Bayes estimator for the disturbance covariance matrix

$$\hat{\Psi} = \frac{\hat{U}}{N + m + \nu - 2p - 2}$$

where

$$\hat{U} \ \ = \ \ (X - \hat{F}\hat{\Lambda}')'(X - \hat{F}\hat{\Lambda}') + (\hat{\Lambda} - \Lambda_0)H(\hat{\Lambda} - \Lambda_0)' + \hat{B}.$$

We can consider $\hat{\Psi}$ as a weighted average of the three terms in $\hat{U}$. The scalar values associated with the terms are N, m, and $\nu - 2p - 2$ respectively. Because we consider the first and third terms as representing the current and training data, we equate $\nu - 2p - 2$ with n to obtain

$$\nu = n + 2p + 2.$$

The mean of the prior distribution for the disturbance covariance matrix is

$$E(\Psi) = \frac{B}{\nu - 2p - 2} \ ,$$

and since $B = b_0 I_p$ the mean of any diagonal element is

$$E(\Psi_{ii}) = \frac{b_0}{\nu - 2p - 2}, \quad i = 1, \ldots, p.$$

From the classical factor analysis model we have

$$\Sigma = \Lambda\Lambda' + \Psi$$

where Σ is the covariance matrix for the observations. Substituting the training sample covariance matrix $\hat{\Sigma}$ and the a priori mean for the factor loadings into the above equation we have

$$\Psi_0 = \hat{\Sigma} - \Lambda_0\Lambda_0'$$

then we take the average of the diagonal elements

$$\frac{1}{p}tr(\Psi_0) = \frac{1}{p}tr(\hat{\Sigma} - \Lambda_0\Lambda_0')$$

as our prior mean for a diagonal element $E(\Psi_{ii})$ of the disturbance covariance matrix. We determine b_0 as

$$b_0 = n\frac{1}{p}tr(\Psi_0).$$

b) We can assess B and ν by purely subjective methods.

Hyperparameters for $p(F|\Phi, m)$

a) We assign R the correlation matrix for factor scores as

$$R = I_m$$

which is the classic orthogonal model.

b) We can use previous experience to assess R.

Hyperparameters for $p(\rho)$

The hyperparameters α and β have the interpretation that $\alpha + \beta - 2$ is the effective prior sample size, and a priori, we believe that for every $\alpha - 1$ times we believe $\rho = b$ we believe there are $\beta - 1$ times $\rho = a$. If for example we expressed no prior beliefs about the value of the parameter ρ then $\alpha = 1$ and $\beta = 1$ can be used which corresponds to a vague or uninformative prior distribution.

a) These are assessed with assistance from the determined correlation structure. For example, if we determined the correlation structure to be intraclass, then $a > -\frac{1}{N-1}$ and $b < 1$. If the structure were first order Markov, then $a > -1$ and $b < 1$.

We assign α and β such that

$$(b-a)\left(\frac{\alpha}{\alpha+\beta}\right)-a$$

is equal to the either the estimated correlation value from the training data.

b) We may also assess by pure subjective prior beliefs of the correlation parameter. We do this because this is the mean of the beta distribution.

D Prior on the Mean

D.1 Extended PS89 Bayesian Factor Analysis

As stated in section 2.4, we can extend the Press and Shigemasu model by assessing a prior distribution for the general mean μ instead of centering the observations with their sample mean (which is the maximum likelihood estimator). The parameters can be estimated in the same fashion as PS89 (EPS89) or as in RP98 (ERP98).

The same factor analysis model applies and is

$$\begin{array}{cccccccccc} (x|\mu,\Lambda,f,m) & = & \mu & + & (I_N \otimes \Lambda) & f & + & \epsilon & . \\ (Np \times 1) & & (Np \times 1) & & (Np \times Nm) & (Nm \times 1) & & (Np \times 1) & \end{array}$$

Using the same likelihood and prior distributions as in the PS89 model but adding the prior distribution on the general mean μ

$$p(\mu|\Psi) \propto |I_N \otimes \Psi|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mu-\mu_0)'(I_N \otimes \psi_0 \Psi)^{-1}(\mu-\mu_0)} \tag{D.1.1}$$

which can be written as the matrix normal distribution

$$p(M|\Psi) \propto |\Psi|^{-\frac{N}{2}} e^{-\frac{1}{2}tr(\psi_0 \Psi)^{-1}(M-M_0)'(M-M_0)}. \tag{D.1.2}$$

We will follow the same conditional modal estimation procedure used in PS89.

First we find the marginal posterior density of the factor scores. Integrating the joint posterior distribution with respect to Ψ, Λ, and M, in that order then applying the large sample result as before, the marginal posterior distribution for the factor scores is approximately matrix-T with mean.

$$\hat{F} \equiv (I_N - XW^{-1}X')^{-1}XW^{-1}\Lambda_0 H. \tag{D.1.3}$$

where

$$W \equiv X'X + B + \Lambda_0 H \Lambda_0' + M_0'(\psi_0)^{-1}M_0,$$

The posterior distribution of the general mean M given the factor scores and the data is approximately matrix-T with mean

$$\hat{M} \equiv \frac{1}{1+\psi_0}\left(I_N - \frac{\psi_0}{1+\psi_0}\hat{F}Q_{\hat{F}}^{-1}\hat{F}'\right)^{-1}[M_0 + \psi_0(X - \hat{F}Q_{\hat{F}}^{-1}(X'\hat{F} + \Lambda_0 H)')],$$

where we define

$$Q_{\hat{F}} \equiv (H + \hat{F}'\hat{F}).$$

In order to obtain the above equation we also assumed that $H = h_0 I_m$ as is usually done.

Taking a closer look at the equation for $\hat{M}$ we see that for large samples

$$\frac{\psi_0}{1+\psi_0}\hat{F}Q_{\hat{F}}^{-1}\hat{F}' \;=\; \frac{\psi_0}{1+\psi_0}\frac{\hat{F}\hat{F}'}{(h_0+N)}.$$

Now recall that F is a matrix of random variables distributed as $N(0,1)$ so each element will be less than 3 in magnitude. Lets consider what happens when we assume an extreme case where all the elements were 3. If we multiply an $N \times m$ matrix of 3's by its transpose, the resulting $N \times N$ matrix has elements that are $m3^2 = 9m$. We now have

$$\frac{\psi_0}{1+\psi_0}\hat{F}\frac{Q_{\hat{F}}^{-1}}{2}\hat{F}' \;=\; \frac{\psi_0}{1+\psi_0}\frac{9m}{h_0+N}J_N$$

where J_N is an $N \times N$ matrix of ones. With a large sample size, the above matrix is approximately zero. This is also aided by h_0 being a positive number and ψ_0 being taken as a small number between zero and one. Thus the estimator for the general mean becomes

$$\hat{M} \;\equiv\; \frac{1}{1+\psi_0}[M_0 + \psi_0(X - \hat{F}(\Lambda_0 H Q_{\hat{F}}^{-1})')]. \tag{D.1.4}$$

The posterior distribution of the factor loadings Λ given the factor scores, the general mean, and the data is approximately matrix-T with mean

$$\hat{\Lambda} \;\equiv\; [(X-\hat{M})'\hat{F} + \Lambda_0 H]Q_{\hat{F}}^{-1}. \tag{D.1.5}$$

The posterior distribution of the disturbance covariance matrix Ψ given the factor scores, the general mean, the factor loadings, and the data is approximately distributed as an inverted Wishart with hyperparameter matrix

$$\begin{aligned}\hat{U} &\equiv (X-\hat{M}-\hat{F}\hat{\Lambda}')'(X-\hat{M}-\hat{F}\hat{\Lambda}')+(\hat{\Lambda}-\Lambda_0)H(\hat{\Lambda}-\Lambda_0)' \\ &+ (\hat{M}-M_0)'\psi_0^{-1}(\hat{M}-M_0)+B,\end{aligned}$$

and hyperparameter degrees of freedom $2N+m+\nu$ so that

$$\begin{aligned}\hat{\Psi} &\equiv \frac{\hat{U}}{2N+m+\nu-2p-2} \\ \hat{\Psi}_{mode} &\equiv \frac{\hat{U}}{2N+m+\nu}.\end{aligned} \tag{D.1.6}$$

We can also apply Gibbs sampling and LSO to this model.

D.2 Extended RP98 Bayesian Factor Analysis

For Gibbs sampling, the procedure is to cycle through

$$\begin{aligned}
\bar{\Lambda}_{(i+1)} &\equiv \text{ a random sample from } p(\Lambda|\bar{M}_{(i)}, \bar{F}_{(i)}, \bar{\Psi}_{(i)}, X, m) \\
\bar{\Psi}_{(i+1)} &\equiv \text{ a random sample from } p(\Psi|\bar{M}_{(i)}, \bar{F}_{(i)}, \bar{\Lambda}_{(i+1)}, X, m) \\
\bar{F}_{(i+1)} &\equiv \text{ a random sample from } p(F|\bar{M}_{(i)}, \bar{\Lambda}_{(i+1)}, \bar{\Psi}_{(i+1)}, X, m) \\
\bar{M}_{(i+1)} &\equiv \text{ a random sample from } p(M|\bar{F}_{(i+1)}, \bar{\Lambda}_{(i+1)}, \bar{\Psi}_{(i+1)}, X, m)
\end{aligned}$$

where the posterior conditional distribution of the loadings $(\Lambda|M, F, \Psi, m, X)$ is normally distributed with mean

$$[(X - M)' + \Lambda_0 H](H + F'F)^{-1}$$

and covariance matrix

$$\Psi \otimes (H + F'F)^{-1},$$

the posterior conditional distribution of the disturbance covariance matrix $(\Psi|M, F, \Lambda, m, X)$ has an inverted Wishart distribution with parameter matrix

$$U = (X-M-F\Lambda')'(X-M-F\Lambda')+(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'+(M-M_0)'\psi_0^{-1}(M-M_0)+B$$

and degrees of freedom

$$2N + m + \nu,$$

the posterior conditional distribution of the scores $(F|M, \Lambda, \Psi, m, X)$ is normally distributed with mean

$$(X - M)\Psi^{-1}\Lambda(I_m + \Lambda\Psi^{-1}\Lambda)^{-1}$$

and covariance matrix

$$I_N \otimes (I_m + \Lambda\Psi^{-1}\Lambda),$$

the posterior conditional distribution of the general mean $(M|F, \Lambda, \Psi, m, X)$ is normally distributed with mean

$$\frac{1}{1+\psi_0}[M_0 + \psi_0(X - F\Lambda')]$$

and covariance matrix

$$\Phi \otimes \left(\frac{\psi_0}{1+\psi_0}\Psi\right).$$

As before we will have

$$(\bar{M}_{(1)}, \bar{\Psi}_{(1)}, \bar{F}_{(1)}, \bar{\Lambda}_{(1)})$$

$$\vdots$$

$$(\bar{M}_{(s)}, \bar{\Psi}_{(s)}, \bar{F}_{(s)}, \bar{\Lambda}_{(s)})$$

$$(\bar{M}_{(s+1)}, \bar{\Psi}_{(s+1)}, \bar{F}_{(s+1)}, \bar{\Lambda}_{(s+1)})$$

$$\vdots$$

$$(\bar{M}_{(s+t)}, \bar{\Psi}_{(s+t)}, \bar{F}_{(s+t)}, \bar{\Lambda}_{(s+t)}).$$

The first s random samples called the "burn in" are discarded and the remaining t samples are kept. The means of the remaining random samples

$$\bar{F} = \frac{1}{t}\sum_{k=1}^{t} \bar{F}_{(s+k)}$$

$$\bar{M} = \frac{1}{t}\sum_{k=1}^{t} \bar{M}_{(s+k)}$$

$$\bar{\Lambda} = \frac{1}{t}\sum_{k=1}^{t} \bar{\Lambda}_{(s+k)}$$

$$\bar{\Psi} = \frac{1}{t}\sum_{k=1}^{t} \bar{\Psi}_{(s+k)}$$

are used as the posterior estimates of the parameters.

For LSO estimation of the parameters, we start with an initial value for $\tilde{M}$ and $\tilde{F}$, say $\tilde{F}_{(0)}$ and $\tilde{M}_{(0)}$ then cycle through

$$\begin{aligned}
\tilde{\Lambda}_{(i+1)} &\equiv [(X - M_{(i)})'\tilde{F}_{(i)} + \Lambda_0 H](H + \tilde{F}'_{(i)}\tilde{F}_{(i)})^{-1} \\
\tilde{\Psi}_{(i+1)} &\equiv [(X - M_{(i)} - \tilde{F}_{(i)}\tilde{\Lambda}'_{(i+1)})'(X - M_{(i)} - \tilde{F}_{(i)}\tilde{\Lambda}'_{(i+1)}) \\
&\quad +(\tilde{\Lambda}_{(i+1)} - \Lambda_0)H(\tilde{\Lambda}_{(i+1)} - \Lambda_0)'
\end{aligned}$$

$$
\begin{aligned}
& +(\tilde{M}_{(i)} - M_0)'\psi_0^{-1}(\tilde{M}_{(i)} - M_0) + B]/(2N + m + \nu) \\
\tilde{F}_{(i+1)} &\equiv (X - M_{(i)})\tilde{\Psi}_{(i+1)}^{-1}\tilde{\Lambda}_{(i+1)}(I_m + \tilde{\Lambda}'_{(i+1)}\tilde{\Psi}_{(i+1)}^{-1}\tilde{\Lambda}_{(i+1)})^{-1} \\
\tilde{M}_{(i+1)} &\equiv \frac{\psi_0}{1+\psi_0}[M_0 + \psi_0(X - \tilde{F}_{(i+1)}\tilde{\Lambda}'_{(i+1)})].
\end{aligned}
$$

until convergence is reached.

D.3 Extended Correlated Bayesian Factor Analysis

Here is a description of CBFA that is extended to have a prior distribution placed on it instead of estimating the mean by the sample mean and centering the data.

The factor analysis model is

$$
\underset{(Np \times 1)}{(x|\mu, m, \Lambda, f)} = \underset{(Np \times 1)}{\mu} + \underset{(Np \times Nm)}{(I_N \otimes \Lambda)} \underset{(Nm \times 1)}{f} + \underset{(Np \times 1)}{\epsilon}, \tag{D.3.1}
$$

the likelihood for the observation vector is

$$
p(x|\mu, \Omega, m, f, \Lambda) = (2\pi)^{-\frac{Np}{2}}|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}[x-\mu-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-\mu-(I_N\otimes\Lambda)f]}, \quad \Omega > 0, \tag{D.3.2}
$$

where we use the same definitions as before.

We use natural conjugate prior distributions to represent our uncertainty about the parameters and assume that the joint prior distribution for the unknown parameters is given by

$$p(\mu,\Omega,m,f,\lambda) = p(\mu|\Omega)p(\Omega)p(m)p(f|m)p(\lambda|m), \tag{D.3.3}$$

where $p(\Omega)$, $p(m)$, $p(f|m)$, $p(\lambda|m)$ are the same as before and

$$p(\mu|\Omega) \quad = \quad (2\pi)^{-\frac{Np}{2}}|\omega_0\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mu-\mu_0)'(\omega_0\Omega)^{-1}(\mu-\mu_0)}. \tag{D.3.4}$$

By Bayes' rule, the joint posterior distribution for the unknown parameters of interest is given by

$$\begin{aligned}
p(\mu,\Omega,m,f,\lambda|x) &\propto p(\mu,\Omega,m,f,\lambda)p(x|\mu,\Omega,m,f,\Lambda) \\
&\propto p(\mu|\Omega)p(\Omega)p(\lambda|m)p(f|m)p(m)p(x|\mu,\Omega,m,f,\Lambda) \\
&\propto |\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mu-\mu_0)'(\omega_0\Omega)^{-1}(\mu-\mu_0)}|\Omega|^{-\frac{\nu}{2}}e^{-\frac{1}{2}tr\Omega^{-1}A} \\
&\cdot (2\pi)^{-\frac{Nm}{2}}|\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}(f-f_0)'\Theta^{-1}(f-f_0)} \\
&\cdot (2\pi)^{-\frac{pm}{2}}|\Delta|^{-\frac{1}{2}}e^{-\frac{1}{2}(\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)} \\
&\cdot p(m)|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}[x-\mu-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-\mu-(I_N\otimes\Lambda)f]} \\
&\propto p(m)(2\pi)^{-\frac{(N+p)m}{2}}|\Omega|^{-\frac{(\nu+2)}{2}}e^{-\frac{1}{2}(\mu-\mu_0)'(\omega_0\Omega)^{-1}(\mu-\mu_0)}e^{-\frac{1}{2}tr\Omega^{-1}A} \\
&\cdot |\Theta|^{-\frac{1}{2}}e^{-\frac{1}{2}f'\Theta^{-1}f}|\Delta|^{-\frac{1}{2}}e^{-\frac{1}{2}(\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)} \\
&\cdot e^{-\frac{1}{2}[x-\mu-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-\mu-(I_N\otimes\Lambda)f]}
\end{aligned} \tag{D.3.5}$$

Conditional Posterior Densities

Here we find the necessary posterior conditional densities. We find that the conditional posterior density for the parameters in the same fashion to be

$$
\begin{aligned}
p(\mu|\Omega,m,f,\lambda,x) &\propto e^{-\frac{1}{2}(\mu-\tilde{\mu})'(\frac{\omega_0}{1+\omega_0}\Omega)^{-1}(\mu-\tilde{\mu})} \\
p(\Omega|\mu,m,f,\lambda,x) &\propto |\Omega|^{-\frac{(\nu+2)}{2}}e^{-\frac{1}{2}tr\Omega^{-1}U} \\
p(f|\mu,\Omega,m,\lambda,x) &\propto e^{-\frac{1}{2}(f-\tilde{f})'\left[\Theta^{-1}+(I_N\otimes\Lambda)'\Omega^{-1}(I_N\otimes\Lambda)\right](f-\tilde{f})} \\
p(\lambda|\mu,\Omega,f,x) &\propto e^{-\frac{1}{2}\gamma} \\
p(m|\mu,\Omega,f,\lambda,x) &\propto p(m)(2\pi)^{-\frac{(N+p)m}{2}}|\Omega|^{-\frac{1}{2}}|\Theta|^{-\frac{1}{2}}|\Delta|^{-\frac{1}{2}}e^{-\frac{1}{2}\tau}
\end{aligned}
$$

where we have defined

$$
\begin{aligned}
\tilde{\mu} &= \frac{1}{1+\omega_0}[\mu_0+\omega_0(x-(I_N\otimes\Lambda)f)] \\
U &= [x-\mu-(I_N\otimes\Lambda)f][x-\mu-(I_N\otimes\Lambda)f]'+(\mu-\mu_0)\omega_0^{-1}(\mu-\mu_0)'+A \\
\tilde{f} &= \left[\Theta^{-1}+(I_N\otimes\Lambda)'\Omega^{-1}(I_N\otimes\Lambda)\right]^{-1}(I_N\otimes\Lambda)'\Omega^{-1}(x-\mu) \\
\gamma &= (\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0)+[x-\mu-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-\mu-(I_N\otimes\Lambda)f] \\
\tau &= [x-\mu-(I_N\otimes\Lambda)f]'\Omega^{-1}[x-\mu-(I_N\otimes\Lambda)f]+f'\Theta^{-1}f \\
&+ (\lambda-\lambda_0)'\Delta^{-1}(\lambda-\lambda_0).
\end{aligned}
$$

The conditional posterior density of the mean given the error covariance matrix, the number of factors, the factor scores, the factor loadings, and the data is normal.The conditional posterior density of the error covariance matrix given the mean of the observations, the number of factors, the factor scores, the factor loadings, and the data is an inverted Wishart. The factor scores given the mean of the observations, the error covariance matrix, the number of factors, the factor loadings, and the data follows a normal distribution. The factor scores given the mean of the observations, the error covariance matrix, the number of factors, the factor loadings, and the data follows a normal distribution.

The conditional posterior distribution for the number of factors is not a recognizable distribution.

Gibbs Sampling Estimation

For Gibbs estimation of the posterior, we start with initial values for μ, Ω, m, f, and λ say $\bar{\mu}_{(0)}$, $\bar{\Omega}_{(0)}$, $\bar{m}_{(0)}$, $\bar{f}_{(0)}$, and $\bar{\lambda}_{(0)}$.

Then for a given number of factors $m = \bar{m}_{(i)}$ cycle through

$$\begin{aligned}
\bar{\mu}_{(i+1)} &\equiv \text{ a random sample from } p(\mu|\bar{\Omega}_{(i)}, \bar{m}_{(i)}, \bar{f}_{(i)}, \bar{\lambda}_{(i)}, x) \\
\bar{\Omega}_{(i+1)} &\equiv \text{ a random sample from } p(\Omega|\bar{\mu}_{(i+1)}, \bar{m}_{(i)}, \bar{f}_{(i)}, \bar{\lambda}_{(i)}, x) \\
\bar{f}_{(i+1)} &\equiv \text{ a random sample from } p(f|\bar{\mu}_{(i+1)}, \bar{\Omega}_{(i+1)}, \bar{m}_{(i)}, \bar{\lambda}_{(i)}, x) \\
\bar{\lambda}_{(i+1)} &\equiv \text{ a random sample from } p(\lambda|\bar{\mu}_{(i+1)}, \bar{\Omega}_{(i+1)}, \bar{m}_{(i)}, \bar{f}_{(i+1)}, x).
\end{aligned}$$

which is the Gibbs sampling algorithm.

For the given number of factors $m = \bar{m}_{(i)}$ we have the sequence

$$\begin{gathered}
(\bar{\mu}_{(1)}, \bar{\lambda}_{(1)}, \bar{\Omega}_{(1)}, \bar{f}_{(1)}) \\
\vdots \\
(\bar{\mu}_{(s)}, \bar{\lambda}_{(s)}, \bar{\Omega}_{(s)}, \bar{f}_{(s)}) \\
(\bar{\mu}_{(s+1)}, \bar{\lambda}_{(s+1)}, \bar{\Omega}_{(s+1)}, \bar{f}_{(s+1)}) \\
\vdots \\
(\bar{\mu}_{(s+t)}, \bar{\lambda}_{(s+t)}, \bar{\Omega}_{(s+t)}, \bar{f}_{(s+t)})
\end{gathered}$$

The first s random samples called the “burn in” are discarded and the remaining t samples are kept. The means of the remaining random samples

$$\bar{\mu} = \frac{1}{t} \sum_{k=1}^{t} \bar{\mu}_{(s+k)} \tag{D.3.6}$$

$$\bar{\Omega} = \frac{1}{t} \sum_{k=1}^{t} \bar{\Omega}_{(s+k)} \tag{D.3.7}$$

$$\bar{f} = \frac{1}{t} \sum_{k=1}^{t} \bar{f}_{(s+k)} \tag{D.3.8}$$

$$\bar{\lambda} = \frac{1}{t} \sum_{k=1}^{t} \bar{\lambda}_{(s+k)} \tag{D.3.9}$$

are used as the posterior mean estimates of the parameters given the number of factors $m = \bar{m}_{(i)}$. We do this for each value of m, then find the value of the number of factors $m = \bar{m}$ that makes the posterior conditional distribution for the number of factors $p(m|\bar{\mu}, \bar{\Omega}, \bar{f}, \bar{\lambda}, x)$ a maximum given the corresponding estimates of the other parameters. This is the same as finding the value for the number of factors that gives the largest conditional posterior odds ratio. We will have $(\bar{m}, \bar{\mu}, \bar{\Omega}, \bar{f}, \bar{\lambda})$ as our posterior estimates of the unknown parameters where $(\bar{\mu}, \bar{\Omega}, \bar{f}, \bar{\lambda})$ are the estimates conditional on $m = \bar{m}$.

As before, the posterior conditional distribution for the factor loadings λ, $p(\lambda|\mu, \Omega, m, f, x)$, the terms in the exponent do not combine nicely to form a well known and recognizable distribution.

References

[1] Alexander Basilevsky. *Statistical Factor Analysis and Related Methods.* John Wiley and Sons Inc., New York, 1994.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[3] J. Dongarra et. al. LAPACK routine (version 2.0) – Univ. of Tennessee, Univ. of California Berkeley, NAG Ltd., Courant Institute, Argonne National Lab, and Rice University March 31, 1993.

[4] A. E. Gelfand and A. F. M Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[5] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984.

[6] W. Gilks and P. Wild. Adaptive rejection sapling for Gibbs sampling. *Applied Statistics*, 41:337–348, 1992.

[7] James D. Hamilton. *Time Series Analysis.* Princeton University Press, Princeton, New Jersey, 1994.

[8] Kentaro Hayashi. *The Press-Shigemasu Bayesian Factor Analysis Model With Estimated Hyperparameters.* PhD thesis, University of North Carolina, Chapel Hill, 1997.

[9] John Imbrie and Nilva Kipp. A new micropaleontological method for quantitative paleoclimatology: Application to a late pleistocene caribbean core. In *The Late Cenozoic Glacial Ages*, chapter 5. Yale University Press, 1971.

[10] IMSL. MATH & STAT/ LIBRARY user's manual version 1.0, 1987.

[11] H. F. Kaiser. The application of electronic computers to factor analysis. *Education and Psychological Measurement*, 20:141–151, 1960.

[12] William Kennedy and James Gentle. *Statistical Computing.* Marcel Dekker, Inc., New York, 1980.

[13] Samuel Kotz and Norman Johnson, editors. *Encyclopedia of Statistical Science*, volume 5. John Wiley and Sons, Inc., New York, 1985. pages 326–333.

[14] D. N. Lawley. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60:64–82, 1940.

[15] Sang Eun Lee. *Robustness of Bayesian Factor Analysis Estimates.* PhD thesis, University of California, Riverside, December 1994.

[16] Sang Eun Lee and S. James Press. Robustness of Bayesian factor analysis estimates. *Communications in Statistics – Theory And Methods*, 27(8), 1998.

[17] D. V. Lindley and A. F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, 34(1), 1972.

[18] Anthony O'Hagen. *Kendalls' Advanced Theory of Statistics, Volume 2B Bayesian Inference.* John Wiley and Sons Inc., New York, 1994.

[19] S. J. Press. Matrix intraclass covariance matricies with applications in agriculture. Technical Report No. 49, Department of Statistics, University of California, Riverside, February 1979.

[20] S. J. Press and K. Shigemasu. Bayesian inference in factor analysis. In *Contributions to Probability and Statistics*, chapter 15. Springer-Verlag, 1989.

[21] S. J. Press and K. Shigemasu. Posterior distribution for the number of factors. Technical Report No. 208, Department of Statistics, University of California, Riverside, April 1994.

[22] S. J. Press and K. Shigemasu. Bayesian inference in factor analysis-Revised. Technical Report No. 243, Department of Statistics, University of California, Riverside, May 1997.

[23] S. James Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference.* Robert E. Krieger Publishing Company, Malabar, Florida, 1982.

[24] S. James Press. *Bayesian Statistics: Principles, Models, and Applications.* John Wiley and Sons, New York, 1989.

[25] Daniel B. Rowe and S. James Press. Gibbs sampling and hill climbing in Bayesian factor analysis. Technical Report No. 255, Department of Statistics, University of California, Riverside, May 1998.

[26] Donald Rubin and Dorothy Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.