ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES AND TECHNOLOGY
DEPARTMENT OF STATISTICS

BAYESIAN FACTOR ANALYSIS

Author:
Nikolaos Fliris

Supervisor:
Panagiotis Papastamoulis

M.Sc. Thesis

Submitted to the Department of Statistics of
Athens University of Economics and Business
in partial fulfilment of the requirements for the degree of
Master of Science in Statistics

Athens,Greece
October 1, 2023

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕ ΜΠΕΥΖΙΑΝΕΣ ΜΕΘΟΔΟΥΣ

Συγγραφέας:
Νικόλαος Φλήρης

Επιβλέπων:
Παναγιώτης Παπασταμούλης

Μεταπτυχιακή Διατριβή

Που υποβλήθηκε στο Τμήμα Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών ως μέρος των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα,Ελλάδα
October 1, 2023

Absract

Factor Analysis is a multivariate statistical technique mainly used to characterize the dependence structure of observed correlated variables using unobserved variables known as factors while at the same time providing dimensionality reduction. The purpose of this thesis is to describe the factor analysis model as well as its assumptions and its characteristics. Moreover, it explains the identification issue of the parameters of the Bayesian factor analysis model, namely the uniqueness problem of the variance covariance matrix of the idiosyncratic errors and the identification issue regarding the factor loadings matrix originated by orthogonal ambiguities and label switching. Furthermore, some well-known Bayesian methods for the production of posterior samples for the parameters of interest of the factor analysis model are described as well as some Bayesian methods solving the identification issues of the parameters of a Bayesian factor analysis model. Additionally, the efficiency of these methods will be evaluated in synthetic data sets and two real data sets. In particular, the two real data sets consists of the Humor Style Questionnaire based on the humor model of R. A. Martin, Puhlik-Doris, Larsen, Gray, and Weir (2003) and the Big Five Personality test, one of the most well-known data sets in the field of psychology.

Περίληψη

Η ανάλυση παραγόντων είναι μια πολυμεταβλητή στατιστική τεχνική που χρησιμοποιείται κυρίως για να χαρακτηρίσει τη δομή εξάρτησης των παρατηρούμενων συσχετιζόμενων μεταβλητών χρησιμοποιώντας μη παρατηρούμενες μεταβλητές γνωστές ως παράγοντες, ενώ ταυτόχρονα παρέχει μείωση των διαστάσεων. Σκοπός της παρούσας διατριβής είναι να περιγράψει το μοντέλο της παραγοντικής ανάλυσης καθώς και τις υποθέσεις και τα χαρακτηριστικά του. Επιπλέον, εξηγεί το ζήτημα της ταυτοποίησης των παραμέτρων ενός Μπεϋζιανού (Bayes) μοντέλου ανάλυσης παραγόντων, συγκεκριμένα, το πρόβλημα της μοναδικότητας του πίνακα διακυμάνσεων συνδιακυμανσεων των ιδιότυπων σφαλμάτων και το πρόβλημα ταυτοποίησης που αφορά τον πίνακα φορτώσεων παραγόντων. Επιπρόσθετα, περιγράφονται ορισμένες γνωστές Μπεϋζιανές μέθοδοι για προσομοίωση δειγμάτων από την εκ των υστέρων κατανομή που αφορούν τις παραμέτρους ενδιαφέροντος του μοντέλου ανάλυσης παραγόντων, καθώς και ορισμένες Μπεϋζιανές μέθοδοι που επιλύουν τα ζητήματα ταυτοποίησης των παραμέτρων ενός Μπεϋζιανού μοντέλου ανάλυσης παραγόντων. Επιπλέον, η αποτελεσματικότητα των μεθόδων αυτών αξιολογείται σε συνθετικά και σε δύο πραγματικά σύνολα δεδομένων. Συγκεκριμένα, τα δύο πραγματικά σύνολα δεδομένων αποτελούνται από το ερωτηματολόγιο Humor Style Questionnaire που βασίζεται στο μοντέλο χιούμορ των R. A. Martin et al. (2003) και το Big Five Personality Test, ένα από τα πιο γνωστά σύνολα δεδομένων στον τομέα της ψυχολογίας.

Outline of the Thesis

The first chapter is an introductory one, where the basic factor analysis model is described along with its characteristics and its assumptions. Furthermore, the maximum likelihood estimation of the parameters of the factor analysis model is described. Additionally, the Estimation Maximization algorithm for the estimation of the parameters of the factor analysis model is presented. Moreover, the prior specifications and assumptions of two simple Bayes factor models are described. The identifiability problem of the parameters of the factor analysis model is described, as well as some well-know methods of classical and Bayesian statistic solving it. Finally, the issue of the selection of the number of factors is discussed as well as some methods in order to identify the appropriate number of factors in a factor analysis model.

The second chapter attempts to describe how each method generates the MCMC samples, the prior specification and the posterior distribution of each method producing MCMC samples, the technique each method employs in order to solve identification issues and, finally, how the algorithm of each method operates. In particular, how the BEFA, Positive Lower Triangular, Dirichlet-Laplace Shrinkage, Multiplicative Gamma Process Shrinkage models produce the MCMC samples, their prior specifications and their posterior distributions. Additionally, it is explained how the BEFA, MatchAlign, RSP exact, RSP full simulation Annealing, RSP partial simulation Annealing,WOP and OP methods solve identification issues.

In the third chapter a comparison of the efficiency of the different models described in the second chapter will be made with the employment of synthetic data sets. Four different domains have been used so that the comparison can be made. The first domain is about how much time each method requires in order to solve rotational ambiguities and label switching. The second domain concern the number of factors selected by the model in comparison with the true number of factors. The third domain analyses how many variables are allocated correctly. Finally, the forth domain presents a metric evaluating the performance of solving rotational ambiguities and label switching.

The forth chapter compares the efficiency of the models described in the second chapter. In order to achieve this two real data sets will be employed. The first one is the Humor Style Questionnaire based on the four humor type model suggested by R. A. Martin et al. (2003) and the second one is the Big Five personality test suggested by Goldberg (1992). The comparison will be made in the same four domains as in chapter three.

Finally, in the fifth chapter a brief discussion of the conclusions and propositions for future work will be made.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Professor Panagiotis Papastamoulis for his continuous support and encouragement. I would not be able to finish my thesis without his enthusiasm, patience, motivation and most of all, his immense knowledge and rich experience.

Secondly, I would like to express my sincere thanks to all the staff at Department of Statistics for all of their support and help through these years.

Last but not least, I would like to thank my family for their unconditional love and support.

Contents

List of Figures

List of Tables

List of Acronyms

MCMC	Markov chain Monte Carlo
EM	Expectation Maximization Algorithm
Anova	Analysis of Variance
PLT	Positive Lower Triangular
RSP	Rotation Sign permutation
WOP	Weighted Orthogonal Procrustes
OP	Orthogonal Procrustes
BEFA	Bayesian Exploratory Factor Analysis
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
ICOMP	Informational Complexity Measurement
CV	Cross Validation
RLME	Reference Loading Matrix Estimation
SP	Signed Permutation
GL	Global Local Mixture of Gaussians
M-H	Metropolis Hastings
DSS	Decoupling Shrinkage and Selection
DSSFA	Decoupling Shrinkage and Selection for Factor Analysis
N	Normal distribution
N_p	p-dimensional normal distribution
U	Uniform distribution
G	Gamma distribution
IG	Inverse-Gamma distribution
Exp	Exponential distribution
DE	Double Exponential or Laplace distribution
Dir	Dirichlet distribution
$IGaussian$	Inverse-Gaussian distribution
$gIGaussian$	Generalized Inverse-Gaussian distribution
$Beta$	Beta distribution
IW	Inverse-Wishart distribution
W	Wishart distribution
TN	Truncated Normal distribution
$Poisson$	Poisson distribution

Chapter 1

Introduction

1.1 Factor Analysis

Factor Analysis is a multivariate statistical technique which is mostly used in social, behavioural and medical studies. Factor Analysis is mainly used to characterize the dependence structure of observed correlated variables using unobserved variables known as factors while providing dimensionality reduction. This is achieved via decompositioning the $p \times p$ covariate matrix of the data Ω as $\Lambda\Lambda^T + \Sigma$ where Λ is a $p \times q$ factor loading matrix with q (number of factors) $<<$ p (number of variables) and Σ is a $p \times p$ diagonal matrix with non-negative diagonal entries.

Before we dive in the factor model, we must clarify some things and have a brief historic overview of the factor model. First there are two different types of factor analysis. The explanatory factor analysis which is data-driven and used when we do not have a-priori knowledge about the number of factors or the relationship patterns of the observed variables and the second type of factor analysis which is called confirmatory factor analysis being theory-driven and used when we have some knowledge about the data and we wish to test whether the data can fit a hypothesized measurement model. This hypothesized model is developed on theory and knowledge or previous analytic research. Confirmatory factor analysis was first developed by Jöreskog (1969). Our main focus will be on explanatory factor analysis, however, in my opinion, we had to make the distinction between those two before we begin.

Conclusively for the historic overview, factor analysis is one of the oldest multivariate methods, but its creation lies outside of the statistical field. The

beginning of factor analysis can be found in the last century in the field of psychology when researchers tried to measure human intelligence. The origins of factor analysis are usually attributed to Spearman (1904). However some researchers support that the key idea was already present in Galton (1889). Thurstone recommended the desirability of acquiring solutions with simple structure (Thurstone, 1935). Moreover, Kendall (1950) set some of the foundations for the blooming of the factor model. In the next subsection of our thesis we will explain the mathematical formulas of the factor analysis model.

1.1.1 Factor analysis model

The typical basic q-factor model is expressed as a linear combination of a latent vector of factors and has the following form

$$Y_i = \mu + \Lambda F_i + e_i \text{ for } i = 1, 2, \ldots, n \tag{1.1}$$

in particular, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ denotes the i-th observation of a random sample of p dimensional observations with $Y_i \in R^p$. q denotes the fixed number of factors where it is obvious that $q > 0$. Additional in order to achieve dimension reduction and for identification reasons, discussed latter in this thesis, $q < p$. $N_p(\mu, \Sigma)$ denotes a p-dimensional normal distribution with covariance matrix Σ and mean $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$. The matrix of factor loadings is denoted as $\Lambda = (\lambda_{rj})$ a $p \times q$ matrix. $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ denotes the marginal mean of Y_i. $F_i = (F_{i1}, F_{i2}, \ldots, F_{iq})^T$ is the unobserved vector of factors. The factors are uncorrelated and distributed as

$$F_i \sim N_q(0_q, \Phi_q), i = 1, \ldots, n. \tag{1.2}$$

The idiosyncratic errors e_i are independent and normally distributed as follows

$$e_i \sim N_q(0_p, \Sigma) \text{ where } \Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2). \tag{1.3}$$

Furthermore, the idiosyncratic errors are independent from the unobserved vector of factors F_i.

The conditional distribution of Y_i on F_i is the following

$$Y_i | F_i \sim N_p(\mu + \Lambda F_i, \Sigma), \ i = 1, \ldots, n. \tag{1.4}$$

For simplicity reasons, for the rest of this thesis we can assume that our data have been standardised before our research meaning that they have

zero means and unit variances. Additionally, the variance covariance matrix of the factors is equal to the identity matrix. This corresponds to assuming independence among all factors and conditional independence among the manifest variables given the factor scores.

Therefore, the marginal distribution of Y_i is the following

$$Y_i \sim N_p(0, \Omega) \text{ where } \Omega = \Lambda\Lambda^T + \Sigma, \ i = 1, \ldots, n. \tag{1.5}$$

The decomposition of $\Omega = \Lambda\Lambda^T + \Sigma$ indicates that the association between the observed variables can be fully explained through the unobserved variables through the quantity $\Lambda\Lambda^T$.

Maximum Likelihood of Factor Analysis Model

Assuming that the data comes from a p-dimensional multivariate Normal distribution $y \sim N_p(\mu, \Omega)$ and supposing that $Y = (y_1, \ldots, y_n)$ is a sample and if and only if $n > p$, the Likelihood function can be calculated by following the notation used by Cao (2010) as follows:

$$L(\mu, \Omega|Y) = f(y_1, \ldots, y_n|\mu, \Omega) = \frac{1}{|2\pi\Omega|^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\{-\frac{1}{2}(y_i - \mu)^T\Omega^{-1}(y_i - \mu)\}$$

the log-Likelihood will be

$$\log(L(\mu, \Omega|Y)) = l(\mu, \Omega|Y) = -\frac{n}{2}\log|2\pi\Omega| - \frac{1}{2}\sum_{i=1}^{n}\{(y_i - \mu)^T\Omega^{-1}(y_i - \mu)\}.$$

With some calculations, it can be proved that the maximum likelihood estimation for μ is $\bar{y}$. Therefore, the log-likelihood function can be rewritten as

$$l(\bar{y}, \Omega|Y) = -\frac{n}{2}\log|2\pi\Omega| - \frac{1}{2}trace\{n\Omega^{-1}S\}$$
$$\text{where } S = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^T(y_i - \bar{y}).$$

After calculating, it can be proved that that the maximum likelihood estimation for Ω is S. Ω has to be substituted with its decomposition $\Omega = \Lambda\Lambda^T + \Sigma$ in order to maximise Λ and Σ and obtain their maximum likelihood estimations

$$l(\bar{y}, \Lambda, \Sigma|Y) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Lambda\Lambda^T + \Sigma| - \frac{1}{2}trace\{(\Lambda\Lambda^T + \Sigma)^{-1}S\}.$$

However, because of non identification issues, in order to proceed with the calculations, a constraint requiring that $\Lambda^T\Sigma\Lambda$ be diagonal needs to be added (Cao, 2010).

Minimizing $l(\bar{y}, \Lambda, \Sigma|Y)$ is the same as maximizing

$$\log|\Lambda\Lambda^T + \Sigma| + trace\{S^{\frac{1}{2}}(\Lambda\Lambda^T + \Sigma)^{-1}S^{\frac{1}{2}}\}.$$

For simplicity reasons, let's denote $(\Lambda\Lambda^T + \Sigma)^{-1}$ as B. Then, the previous equation can be written as

$$\log|B^{-1}| + trace\{S^{\frac{1}{2}}BS^{\frac{1}{2}}\} = trace\{S^{\frac{1}{2}}BS^{\frac{1}{2}}\} - \log|S^{\frac{1}{2}}BS^{\frac{1}{2}}| + \log|S|$$

where $trace(S^{\frac{1}{2}}BS^{\frac{1}{2}})$ is the summation of the eigenvalues of $S^{\frac{1}{2}}BS^{\frac{1}{2}}$, denoted as $\sum_{i=1}^{n}\gamma_i$ and $\log|S^{\frac{1}{2}}BS^{\frac{1}{2}}|$ is the logarithmic product of the eigenvalues, denoted as $\prod_{i=1}^{n}\gamma_i$. In order to continue with the calculations, we need to calculate the derivatives with respect to the eigenvalues γ_i and equate them to zero. Achieving that, a set of equations occurs (Cao, 2010)

$$1 - \frac{1}{\gamma_i} = 0 \ ,i = 1, \ldots, p$$

so $\hat{\gamma_i} = 1$. Consequently, the log likelihood function is maximized at (Λ, Σ) fulfilling

$$S^{\frac{1}{2}}(\hat{\Lambda}\hat{\Lambda}^T + \hat{\Sigma})^{-1}S^{\frac{1}{2}} = I.$$

Since $S^{\frac{1}{2}}(\Lambda\Lambda^T + \Sigma)^{-1}S^{\frac{1}{2}}$ is symmetrical and because $S = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Sigma}$ and by denoting $D = \hat{\Lambda}^T\hat{\Sigma}^{-1}\hat{\Lambda}$ (where D is some diagonal matrix) the previous equation can be rewritten as

$$\hat{\Lambda}(I + D) = S\hat{\Sigma}^{-1}\hat{L}.$$

Consequently, we will have the following equations which we must solve for $\hat{\Lambda}, \hat{\Sigma}$

$$\begin{aligned}\hat{\Lambda}(I + D) &= S\hat{\Sigma}^{-1}\hat{\Lambda}\\ D &= \hat{\Lambda}^T\hat{\Sigma}^{-1}\hat{\Lambda} \text{ diagonal}\\ \hat{\Sigma} &= S - \hat{\Lambda}\hat{\Lambda}^T \text{ diagonal.}\end{aligned}$$

There is not a closed analytical solution for the above equations. So, Lawley and Maxwell (1962) suggested the following iterative procedure

Step 1 Approximate $\hat{\Sigma}$

Step 2 Solve for $\hat{\Lambda}$

Step 3 Use $\hat{\Sigma} = S - \hat{\Lambda}\hat{\Lambda}^T$ to obtain a new approximation for $\hat{\Sigma}$.

EM Estimation of the Maximum Likelihood of Factor Analysis Model

The Maximum Likelihood estimation of a factor analysis model using the Expectation Maximization (EM) algorithm was first proposed by Rubin and Thayer (1982) with the following assumptions : 1) That the data comes from a p-dimentional multivariate Normal distribution $Y \sim N_p(0, \Omega)$. 2) That $Y = (y_1, \ldots, y_n)$ is a sample of the likelihood function. 3) That the unobserved Factors are treated as missing data. Taking the above into account, the maximum likelihood estimation of a factor analysis model using the Expectation Maximization (EM) algorithm can be calculated as follows:

The complete data Likelihood function is

$$L_C(\Lambda, \Sigma) = \left[2\pi \prod_{j=1}^{p} \sigma_j^2\right]^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(y_{ij} - F_i \Lambda_j)^2}{\sigma_j^2}\right] [2\pi]^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} F_i^T F_i\right].$$

The complete data log-Likelihood function is

$$l_c(\Lambda, \Sigma) = \log L_c(\Lambda, \Sigma) = -\frac{n}{2} \sum_{j=1}^{p} \log \sigma_j^2 - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(y_{ij} - F_i \Lambda_j)^2}{\sigma_j^2} - \frac{1}{2} \sum_{i=1}^{n} F_i^T Fi$$

where Λ_j denotes the j-th row of Λ.

E-Step computation

For the calculating of $\mathbb{E}\left[l_c\left(\Lambda, \Sigma | Y, \Lambda_t, \Sigma_t\right)\right]$ it is essential to calculate the expectation of the following sufficient statistics:

$$C_{YY} = \sum_{i=1}^{n} \frac{y_i^T y_i}{n}, \text{ a } p \times p \text{ observed matrix}$$

$$C_{YF} = \sum_{i=1}^{n} \frac{y_i^T F_i}{n}, \text{ a } p \times q \text{ matrix}$$

$$C_{FF} = \sum_{i=1}^{n} \frac{F_i^T F_i}{n}, \text{ a } q \times q \text{ matrix.}$$

Given Σ^2 and Λ the observed variables y_i follow a p-variate normal and the factors F_i a q-variate Normal. Consequently, for given Σ^2 and Λ, the conditional distribution of $F_i|y_i$ is a q-variate Normal with mean δy_i and covariance Δ. Where δ denotes the regression coefficient and Δ the residual covariance matrix. The regression coefficient and the residual covariate matrix can be calculated as

$$\delta = \left(\Sigma_t^2 + \Lambda_t^T \Lambda_t\right)^{-1} \Lambda_t^T$$
$$\Delta = I - \Lambda_t \left(\Sigma_t^2 + \Lambda_t^T \Lambda_t\right)^{-1} \Lambda_t^T.$$

By employing Woodbury's identity (Woodbury, 1949, 1950) $\left(\Sigma_t^2 + \Lambda_t^T \Lambda_t\right)^{-1} = \Sigma_t^{-2} - \left(\Sigma_t^{-2} \Lambda_t^T\right) \left(I + \Lambda_t \Sigma_t^{-2} \Lambda_t^T\right)^{-1} \left(\Lambda_t \Sigma_t^{-2}\right)$, the previous equations can be rewritten as:

$$\begin{aligned} \delta &= \left(\Sigma_t^{-2} - \left(\Sigma_t^{-2} \Lambda_t^T\right) \left(I + \Lambda_t \Sigma_t^{-2} \Lambda_t^T\right)^{-1} \left(\Lambda_t \Sigma_t^{-2}\right)\right) \Lambda_t^T \\ \Delta &= I - \Lambda_t \left(\Sigma_t^{-2} - \left(\Sigma_t^{-2} \Lambda_t^T\right) \left(I + \Lambda_t \Sigma_t^{-2} \Lambda_t^T\right)^{-1} \left(\Lambda_t \Sigma_t^{-2}\right)\right) \Lambda_t^T. \end{aligned} \tag{1.6}$$

Consequently, the conditional expectation of the previously mentioned sufficient statistics, for given Σ^2,Λ and observed variables Y are the following:

$$\begin{aligned}
&\mathbb{E}\left(C_{YY}|Y,\Lambda_t,\Sigma_t\right) = C_{YY} \\
&\mathbb{E}\left(C_{YF}|Y,\Lambda_t,\Sigma_t\right) = C_{YY}\delta \\
&\mathbb{E}\left(C_{FF}|Y,\Lambda_t,\Sigma_t\right) = \delta^T C_{YY}\delta + \Delta.
\end{aligned} \tag{1.7}$$

Consequently, the E-step is characterized by the equations 1.6 1.7 for observed variables Y and present time estimates of Σ_t^2 and Λ_t.

M-Step computation

The succeeding value of the parameters is acquired by maximizing the expected log-Likelihood in the E-Step by employing the expected values of the sufficient statistics as if they were the observed values.

The maximization will give as the following maximum likelihood estimations:

$$\begin{aligned}
\Lambda_{t+1} &= \left(\delta^T C_{YY}\delta + \Delta\right)^{-1}\left(C_{YY}\delta\right)^T \\
\Sigma_{t+1}^2 &= diag\left(C_{YY} - C_{YY}\delta\left(\delta^T C_{YY}\delta + \Delta\right)^{-1}\left(C_{YY}\delta\right)^T\right).
\end{aligned} \tag{1.8}$$

Employing the estimations in 1.8 instead of Λ_t and Σ_t^2 in 1.6 generates the new values of δ and Δ. Then these new values are employed in equations 1.8 to generate new values of Λ_{t+1} and Σ_{t+1}^2, and so on.

1.1.2 Typical Bayesian Factor Analysis Models

In the Bayesian factor analysis models as in any Bayesian models a prior specification for the parameters of interest is required. In this subsection, two simple prior specifications will be illustrated so that the functionality of these factor analysis models can be comprehended. More complicated prior specifications and models will be presented Chapter 2.

Both models assume that the number of factors is known.

Used by Polasek (1997), the first prior specification is more general and its notation will be used as follows:

The rows of the factors loadings matrix Λ will follow a multivariate Normal distribution

$$\Lambda \sim N(\Lambda_\star, H_\star \otimes G_\star).$$

The covariance matrix Ω of the observed variables will follow a Wishart distribution

$$\Omega^{-1} \sim W(\Omega_\star, n_\star)$$

where the scaling matrix parameter $\Omega_\star$ restricts the Wishart distribution to diagonal covariance matrices.

The covariance matrix of the factors will follow a Wishart distribution

$$\Phi^{-1} \sim W(\Phi_\star, v_\star).$$

Taking into consideration the assumption that the idiosyncratic errors follow a normal distribution and the previously mentioned prior specifications, then the factors and the observed variables follow the following distributions

$$F \sim N(0, I \otimes \Phi)$$
$$Y \sim N(\Lambda F, I \otimes \Omega)$$

where $\otimes$ denotes the tensor product.

The joint distribution of the observed variables Y and the parameters $\Theta = (\Lambda, F, \Omega, \Phi)$ is the following

$$P(Y, \Theta) = N(vec(Y)|vec(\Lambda F), I \otimes \Omega) N(vec(F)|0, I \otimes \Phi)$$
$$N(vec(\Lambda)|\Lambda_\star, H_\star \otimes G_\star) W(\Omega^{-1}|\Omega_\star, n_\star) W(\Phi^{-1}|\Phi_\star, v_\star)$$
$$P(Y, \Theta) \propto |I \otimes \Omega|^{\frac{1}{2}} \exp\left(-\frac{1}{2} vec^T(Y - \Lambda F)^T (I \otimes \Omega^{-1}) vec(Y - \Lambda F)\right)$$
$$|I \otimes \Phi|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} vec^T F^T (I \otimes \Phi^{-1}) vec(F)\right) |H_\star \otimes G_\star|^{-\frac{1}{2}}$$
$$\exp\left(-\frac{1}{2} vec^T(\Lambda - \Lambda_{star})(H_\star \otimes G_\star)^{-1} vec(\Lambda - \Lambda_\star)\right)$$
$$|\Phi|^{-\frac{n_\star - p - 1}{2}} \exp\left(-\frac{1}{2} trace(\Omega^{-1}\Omega_\star)\right) |\Phi|^{-\frac{v_\star - q - 1}{2}} \exp\left(-\frac{1}{2} trace(\Phi^{-1}\Phi_\star)\right)$$

where *vec* denotes the linear transformation of a matrix into a vector. This is achieved by piling the columns of the matrix one above the other.

The full conditional distribution of the factors F is

$$P(F|\Theta, Y) = N(F_{\star\star}, D_F)$$
$$\text{where } D_F^{-1} = I \otimes \Phi^{-1} + I \otimes \Lambda^T \Omega^{-1} \Lambda$$
$$\text{and } vec(F_{\star\star}) = D_F vec(\Lambda^T \Omega^{-1} Y).$$

The full conditional distribution of the factor loadings matrix Λ is

$$P(\Lambda|\Theta, Y) = N(\Lambda_{\star\star}, C_{\star\star})$$
$$\text{where } C_{\star\star}^{-1} = H_\star^{-1} \otimes G_\star^{-1} + FF^T \otimes \Omega^{-1}$$
$$\text{and } vec(\Lambda_{\star\star}) = C_{\star\star} \left((H_\star^{-1} \otimes G_\star^{-1}) vec(\Lambda_\star) + (F \otimes \Omega^{-1}) vec(Y) \right)$$
$$vec(\Lambda_{\star\star}) = C_{\star\star} \left(vec(G_\star^{-1} \Lambda_\star H_\star^{-1}) + \Omega^{-1} Y F \right).$$

The full conditional distribution of the covariance matrix Ω

$$P(\Omega^{-1}|\Theta, Y) = W(\Omega_{\star\star}, n_{\star\star} = n_\star + n)$$
$$\text{where } \Omega_{\star\star} = \Omega_\star + (Y - \Lambda F)(Y - \Lambda F)^T.$$

The full conditional distribution of the variance matrix Φ

$$P(\Phi^{-1}|\Theta, Y) = W(\Phi_{\star\star}, v_{\star\star} = v_\star + n)$$
$$\text{where } \Phi_{\star\star} = \Phi_\star + FF^T.$$

The second prior specification is more specific and it was implemented by Papastamoulis (2018, 2020). It was originally used in order to simulate MCMC samples for models of mixtures of factor analyzers. However, by setting the number of clusters to 1 and using the UUU restrictions, this prior specification can be used to simulate MCMC samples for a simple factor analysis model.

The prior distribution of the rows of the factor loadings matrix is the following

$$\Lambda_r \sim N(0, L^2), r = 1, \ldots, p$$

where Λ_r denotes the r-th row of the factor loadings matrix.

The prior distribution of the diagonal elements of the variance matrix(the L matrix is diagonal) of the factor loadings is the following

$$L_l^2 \sim IG(g,h),\ l = 1,2,\ldots,q.$$

The prior distribution of the elements of the variance of the idiosyncratic errors is the following

$$\sigma_r^{-2} \sim G(a,b), r = 1,\ldots,p.$$

The joint probability density function of the model is

$$f(Y,F,\Sigma,\Lambda) = f(Y|F,\Lambda,\Sigma)f(\Sigma)f(\Lambda|L)f(L)f(F).$$

The full conditional posterior distribution of the factor loadings is

$$\Lambda_r|\cdots \sim N_r\left(\left[L^{2^{-1}} + \frac{\sum_{i=1}^n F_i F_i^T}{\sigma_r^2}\right]^{-1}\frac{\sum_{i=1}^n Y_{ir}F_i^T}{\sigma_r^2}, \left[L^{2^{-1}} + \frac{\sum_{i=1}^n F_i F_i^T}{\sigma_r^2}\right]^{-1}\right), r = 1,\ldots,p.$$

The full conditional posterior distribution of the idiosyncratic errors variance is the following

$$\sigma_r^{-2}|\cdots \sim G\left(a+\frac{n}{2}, b+\frac{\sum_{i=1}^n (Y_{ir}-\Lambda_r F_i)^2}{2}\right), r = 1,\ldots,p.$$

The full conditional distribution of the factors F is

$$F_i|\cdots \sim N_q\left((I_q+\Lambda^T\Sigma^{-1}\Lambda^T)^{-1}\Lambda^T\Sigma^{-1}Y_i, (I_q+\Lambda^T\Sigma^{-1}\Lambda^T)^{-1}\right), i = 1,\ldots,n.$$

Finally, in order for a Gibbs sampler to be implemented with the previously mentioned prior specification, the following steps have to be followed

Step 1 Give initial values for $(L^{(0)}\Lambda^{(0)}, \Sigma^{(0)}, F^{(0)})$

Step 2 In each iteration $T = 1,\ldots T$

Step 2-a Update $L^{(t)} \sim \left(L|\Lambda^{(t-1)}\right)$

Step 2-b Update $\Lambda^{(t)} \sim \left(\Lambda|L^{(t)}, \Sigma^{(t-1)}, Y, F^{(t-1)}\right)$

Step 2-c Update $\Sigma^{(t)} \sim (\Sigma|Y)$

Step 2-d Update $F^{(t)} \sim \left(F|Y, \Sigma^{(t)}, \Lambda^{(t)}\right)$

For more information about the Gibbs sample the reader is referred to the appendix of this thesis.

Finally, an important note regarding the later prior specification is that if the stochastic value of the variance matrix of the factor loading matrix is change to a pre-determined value, the prior specification coincides with the prior specification in the MCMCpack ("MCMCpack", n.d.). Therefore the Gibbs sampler of the MCMCpack of R(Team, 2021) is similar with the one described above. The same applies to the posterior distributions of the parameters of interest.

1.1.3 Identifiability problems

The identification problems on the typical factor model occur because of two reasons. The first reason is the identifiability of the variance covariance matrix of the idiosyncratic errors Σ. This problem is also called the uniqueness problem. Lets now examine some well know results than ensure the identifiability of Σ. Assuming than the number of factors in the model are q, then the number of free parameters in the covariance matrix $\Omega = \Lambda\Lambda^T + \Sigma$ is $p + pq - \frac{1}{2}q(q-1)$. In the unconstrained covariance matrix of Y_i the number of free parameters is $\frac{1}{2}p(p+1)$. Consequently, if each measurement (observed variable) is related to at most one latent factor (simple structure) then the number of parameters in the covariance matrix is reduced by

$$\tfrac{1}{2}p(p+1) - \left(p + pq - \tfrac{1}{2}q(q-1)\right) = \tfrac{1}{2}\left((p-q)^2 - (p+q)\right).$$

The above equation is positive if $q < \frac{2p+1-\sqrt{8p+1}}{2}$ the quantity $\varphi(p) = \frac{2p+1-\sqrt{8p+1}}{2}$ is know as the Ledermann bound (Ledermann, 1937). It can be shown that Σ is almost surely unique when $q < \varphi(p)$ (Bekker & ten Berge, 1997). Throughout this thesis, it will be assumed that the number of factor is less than the Ledermann bound. The non identifiability does not necessarily take place when $q > \varphi(p)$. In the isotropic error model (a special case of the typical factor model) if we assume $\Sigma = \sigma^2 I_p$ then we conclude to the probabilistic principal component analysis framework of Tipping and Bishop (1999). In those cases the number of factors can reach $p - 1$.

Anderson and Rubin (1956) proved than in order to achieve identifiability of Σ there must be at least three non-zero elements in each column of the factor loading matrix.

Another reason which leads to the non identifiability of Σ is instances of Heywood cases (Heywood, 1931). Heywood cases occurs when the estimation of one or more of the σ_i $i = 1, \ldots, p$ lies outside of the parameter space. For simplicity reasons that means that one or more $\sigma_i < 0$. If the variance of the idiosyncratic errors is zero $\sigma_i = 0$ that means that the variation of the observed variable in question was wholly explained by the factor. Researchers have found that the main reasons Heywood cases occurs are the following.

1. small sample
2. σ_i true value close to zero
3. false sampling
4. small number of observed variables
5. trying to estimate more factors than the true number of them (Over-estimating the number of factors)

In this thesis, a Bayesian setup will be followed, so this problem is totally avoided. This is done by setting priors for the variance of the idiosyncratic errors with posterior distribution with support on $(0, \infty)$. Those priors are usually members of the inverse Gamma family of distributions. Nevertheless, it is common that the posterior distributions of idiosyncratic variances are multimodal, with one mode lying in areas close to zero (Bartholomew, Knott, & Moustaki, 2011).

An important note is the following: a Bayesian analogue of the Heywood case can occur with the employment of standard improper reference priors for the diagonal elements of the variance covariance matrix of the idiosyncratic errors (J. K. Martin & McDonald, 1975; Ihara & Kano, 1995; Lopes & West, 2004) .

Given identifiability of Σ, the second reason for identifiability issues is related to orthogonal transformations of the matrix of factor loadings. Non-identifiability of the factor loadings matrix is a well known problem. Non-identifiability causes multimodality in the objective function in the frequentist setup. However, Lawley and Maxwell (1962) showed that each mode is equivalently optimal for inference. This means that it suffices to choose a

single mode. A well-known method of choosing one mode is by using orthogonalization methods. Some of the orthogonalization methods are Varimax (Kaiser, 1958) and Quartimax (Neuhaus & Wrigley, 1954)

Varimax Rotation

The Varimax problem (Kaiser, 1958) can be summarized in the following sentence: If there is a $p \times q$ factor loadings matrix Λ, which is the $q \times q$ rotational matrix Φ with the attribute of maximizing the sum of within-factor variances of squared factor loadings of the rotated matrix of factor loadings $\tilde{\Lambda} = \Lambda\Phi$. This optimization problem can by written as

$$\text{Maximize } \frac{1}{4}\sum_{j=1}^{q}\left[\sum_{r=1}^{p}\tilde{\lambda}_{rj}^{4} - \frac{1}{p}\left(\sum_{r=1}^{p}\tilde{\lambda}_{rj}\right)^{2}\right]$$

$$\text{subject to } \Phi^{T}\Phi = I_q$$

$$\text{where } \tilde{\lambda}_{rj} = \sum_{k=1}^{q}\lambda_{rk}\varphi_{kj} \; r = 1, \ldots, p \; j = 1, \ldots, q.$$

Kaiser suggested the increase of the objective function by consecutively rotating pairs of factor in order to solve the Varimax problem.

Before thoroughly explaining the causes of the identifiability issues regarding the matrix of factor loadings, some concepts have to be clarified. A square matrix P is an orthogonal matrix if and only if its inverse equals its transpose $P^T = P^{-1}$. Furthermore, the determinant of any orthogonal matrix is equal to 1 when the represent rotation is proper and equal to -1 when the represent rotation is improper. The decomposition $\Omega = \Lambda\Lambda^T + \Sigma$ is not unique and this has as a consequence that Λ is non-identifiable. To simplify the last sentence, assuming that there is a $q \times q$ orthogonal matrix P, $(PP^T = I)$ then $\Lambda' = \Lambda P$ also satisfies the equation. Its after effect is that the posterior distribution of factor loadings is multimodal leading the sampler to jump from one mode to another during the sampling of the factor loadings, thus preventing the algorithm from converging to a single mode.

In order to tackle the identifiability issues regarding the matrix of factor loadings in the Bayesian setup there are two schools of thought. The first one is ad hoc, it achieves identification in the majority of the occasions but at the

cost of losing flexibility. This is achieved by setting specific prior distribution to some of the factor loadings. Some researchers who follow this school of thought are J. Geweke and Zhou (1996), Aguilar and West (2000), Carneiro, Hansen, and Heckman (2003), Bernardo et al. (2003), Lopes and West (2004), Lucas et al. (2006), Mavridis and Ntzoufras (2014) and Papastamoulis (2018, 2020) and. These researchers used methods relying on a lower triangular specification for the factor loadings matrix about which we will discuss in the next chapter of this thesis. The pioneer of the lower triangular methods was Anderson and Rubin (1956).

The second school of thought is post hoc. These methods do not constrain the parameter space a priori. They rather fix the identification issues post hoc via post processing the output of the Gibbs sampler using algorithms solving rotational ambiguities and label switching problems. The following methods were proposed so that the issue could be solved:

Rotation Sign Permutation post processing algorithm (RSP) was created by Papastamoulis and Ntzoufras (2022a). RSP has three different variants. The RSP exact variant which requires the solution of 2^q(q is the number of factors) assignment problems per MCMC iteration. This version of the algorithm is most efficiently used with relative small number of factors. The second variant is the Fully Simulated Anneling RSP which has the benefit to be the fastest of the three, but with a small penalty in the efficiency of the solutions. This method can be used for any logical number of factors. Finally, the last variance is the Partial Simulated Anneling RSP which is slower than Fully Simulated Anneling, but with an increase in the efficiency of the solutions. This method can also be used for any logical number of factors. We will explain these methods in detail in the next chapter of this thesis.

Weighted Orthogonal Procrustes (WOP) was created by Aßmann, Boysen-Hogrefe, and Pape (2016). This method was created in order to solve the identifiability issues for static and dynamic factor models. This method transforms the unconstrained Gibbs sampler output using a sequence of orthogonal matrices. This transformation is called Orthogonal Procrustes rotations which minimizes the posterior expected loss. Aßmann et al. (2016) defined the unconstrained Gibbs sampler as a Gibbs sampler which we do not impose any constraints on the loading matrix in order to solve the identification issue. These methods will be analysed in the next chapter of this thesis .

MatchAlign was created by Poworoznek, Ferrari, and Dunson (2021). This method can be simplified to three steps. In the first step the method orthog-

onalizes the posterior sample of the factor loadings using Varimax rotation. However, it can use any other orthogonalization procedure . In the second step it selects a reference matrix which is called the pivot matrix. In the final step the algorithm pairs the columns of each posterior matrix of factor loadings with the columns of the pivot matrix. This method can be used for any logical number of factors. This method will be also analysed in the next chapter of this thesis .

Bayesian Exploratory Factor Analysis (BEFA) was created by Conti, Frühwirth-Schnatter, Heckman, and Piatek (2014). This method can be used with continuous and binary data. It is a pioneering method because it makes inference in models with simple dedicated structure and with correlated factors. Conti et al. (2014) define dedicated structure as the factor model where all measurements load onto at most one factor. Moreover, BEFA is an fully Bayesian approach because at the same time produces the posterior distribution of the parameters of interest, chooses the number of factors in the factor model and allocates the observed measurements to factors and to the equivalent factor loadings. Additionally, in order to achieve identification, this method applies classical identifying criteria to the prior distribution of the parameters of the factor model. For identification to by achieved, the method restricts the sampler to stay in regions where only identified models are generated. The proposed realizations by this method have the same likelihood and they are similar up to column and sign switching. To deal with the column switching problem, during each iteration the method switches the position of the non-zero columns so that the top elements appear in increasing order. Afterwards, to solve the sign switching problem Conti et al. (2014) employ a point of reference factor loading in each column,(the one having the highest posterior probability of being non-zero in each column). Subsequently, the signs are changed in every iteration so that it can match with those of the point of reference.

1.1.4 Choice of the number of factors

A major problem in factor analysis under both the Frequentist and Bayesian setup is the choice of the number of factors. That is because even when the true number of factors is small, adding more factors will enhance the fit to the data, making the selection of the number of factors an extremely hard procedure.

One way of solving the issue of selecting the number of factors is viewing the

problem as a model selection one. However, as already mentioned, adding more factors will always make the model have a better fit to the data. In order to estimate the number of factors, a method requiring a trade-off between model complexity and model fit has to be used. Such methods are the following:

Akaike Information Criterion (AIC) created by Akaike (1973)

In order to estimate the correct number of factors using AIC the following steps should be followed.

Step 1 Calculate the Akaike Information Criterion for each model. The calculation of AIC for a factor model with q factors is the following:

$$AIC = n\left(p\log(2\pi) + \log|\hat{\Omega}| + trace(\hat{\Omega}^{-1}S)\right) + 2p(q-1) - \frac{q(q-1)}{2}$$
$$\text{where } \hat{\Omega} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Sigma}.$$

Step 2 Select the model with the minimum AIC and select the number of factors of that model as the correct number of factors.

Bayesian Information Criterion (BIC) created by Schwarz (1978)

In order to estimate the correct number of factors using BIC, these steps should be followed:

Step 1 Calculate the Bayesian Information Criterion for each model. The calculation of BIC for a factor model with q factors is the following:

$$BIC = n\left(p\log(2\pi) + \log|\hat{\Omega}| + trace(\hat{\Omega}^{-1}S)\right)$$
$$+\log\left(n - \frac{(2p+1)}{6} - \frac{2q}{3}\right)p(q-1) - \frac{q(q-1)}{2}$$
$$\text{where } \hat{\Omega} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Sigma}.$$

Step 2 Select the model with the minimum BIC and select the number of factors of that model as the correct number of factors.

Informational Complexity Measurement (ICOMP) created by Bozdogan and Ramirez (1987)

In order to estimate the correct number of factors using ICOMP, these steps should be followed:

Step 1 Calculate the Informational Complexity Measurement for each model. The calculation of ICOMP for a factor model with q factors is the following:

$$ICOMP = n\left(p\log(2\pi) + \log|\hat{\Omega}| + trace(\hat{\Omega}^{-1}S)\right)$$
$$+2(q+1)\left(\frac{p}{2}\log(trace(\frac{\hat{\Sigma}}{p})) - \frac{1}{2}\log|\hat{\Sigma}|\right)$$
$$\text{where } \hat{\Omega} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Sigma}.$$

Step 2 Select the model with the minimum ICOMP and select the number of factors of that model as the correct number of factors.

The previously mentioned Information criteria can be implemented under both the Bayesian and the Frequentist setup in order to estimate the appropriate number of factors. The following methods can be implemented under Bayesian setup.

Lee and Song (2002) proposed a method selecting the appropriate model and consequently the correct number of factors. The method selects the correct model with the Bayes Factor (Berger, 2013). In order to compute the Bayes Factor, Lee and Song created a procedure inspired by Path Sampling (Gelman & Meng, 1998).

The Decoupling Shrinkage and Selection for Factor Analysis (DSSFA) method proposed by Bolfarine, Carvalho, Lopes, and Murray (2022) encapsulates the information in the posterior distribution and acquires illustratable point estimates for the parameters of the factors model. The method was inspired by the Decoupling Shrinkage and Selection (DSS) method proposed Hahn and Carvalho (2015).

Finally, as firstly proposed by Papastamoulis and Ntzoufras (2022a), a more trial and error method of selecting the number of factors when the identification issues are resolved is delineated as follows: Begin fitting factor models from the model with a single factor up to the model with Ledermann bound number of factors. Resolve the identification issues for all the models. Examine which the first model is where at least one of those two conditions is true. The first condition is that a factor has all its factor loadings extremely small (some researchers define as small the factor loadings whose absolute value is less than 0.30). The second condition is that a factor has the smallest absolute value in each row of the factor loadings matrix. After identifying this model, we define this model number of factors as k, then we select the number $k-1$ as the appropriate number of factors.

Under a Frequentist setup the following methods can by implemented:

The Elbow method by Raymond (1966) plots the eigenvalues in a decreasing order in a scree-plot. After that, the number of factors the researcher chooses is equal to the number of eigenvalues before the elbow in the plotted curve.

The Kaiser-Guttman criterion Kaiser (1960) and Guttman (1954) chooses the number of factors as the number of eigenvalues of the covariance matrix above 1.

The out-of-sample prediction error method by Haslbeck and van Bork (2022) calculates the model-implied covariance matrix for each factor model from the model with 1 factor up to the model with p factor, where p is the number of variables. The method acquires p regression models for every factor model. Then, for each observed variable X_i it uses the regression models in order to calculate an estimation of the out-of-sample prediction error using cross-validation. Afterwards, it calculates a single aggregate prediction error for each factor model by calculating the mean prediction error across folds and variables. In the end, the method chooses the number of factors of the model with the smallest out-of-sample prediction error.

For the rest of this thesis, whenever it is required to pre-specify the number of factors in a method the trial and error method proposed by Papastamoulis and Ntzoufras (2022a) will be used.

Chapter 2

Inference and identifiability of Bayesian Factor Analysis models

The main objective of the second chapter of this thesis is to explain the way different methods generate the MCMC samples, the prior and posterior distribution of each method producing MCMC samples, how each method solves the identification issue and, finally, how the algorithm of each method operates.

2.1 Positive Lower Triangular

The Positive Lower Triangular matrix (PLT) model firstly proposed by Anderson and Rubin (1956) is an ad hoc method of resolving the identification issue of the factor loadings matrix by producing the MCMC samples of it with some restrictions in some elements of the matrix. This is achieved by setting specific prior distribution to some of the factor loadings elements and pre-specifying some elements equal to 0. The PLT model can be implemented with any method producing MCMC samples for the parameters of the factor analysis model. In order to generate samples with the PLT constrains, WinBUGS(Lunn, Thomas, Best, & Spiegelhalter, n.d.) or OpenBUGS(Surhone, Tennoe, & Henssonow, 2010) software can be used. Another option would be the programming language R (Team, 2021) using the packages Stan (Carpenter et al., 2017) or mcmcpack ("MCMCpack", n.d.). The previously mentioned constraints are the following:

The factor loading matrix must have the following form

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0 & 0 & \dots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \dots & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{q1} & \lambda_{q2} & \lambda_{q3} & \dots & \lambda_{qq} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \lambda_{p3} & \dots & \lambda_{pq} \end{pmatrix}$$

and the diagonal elements of the matrix must be strictly positive($\lambda_{ii} > 0$).

The factor loading matrix having this lower triangular form ensures that it is of full rank q. J. F. Geweke and Singleton (1980) proved that if the factor loading matrix has rank $r < q$, then there is such a matrix Q about which $\Lambda Q = 0$, $Q^T Q = I$ and for any orthogonal matrix M

$$\Omega = \Lambda\Lambda^T + \Sigma = (\Lambda + MQ^T)(\Lambda + MQ^T) + (\Sigma - \Lambda\Lambda^T).$$

Consequently, the variance covariance matrix Ω is non-identifiable.

An important note about this method is that the choice of the diagonal elements is a key modelling decision which impacts on the model fit and the interpretation of the factors. Carvalho et al. (2008) refers to the diagonal elements of the factor loading matrix as Founders of the Factors. Also, the ordering of the rest of the variables affect the model fit but not to the same extent as the Founders of the Factors.

By restricting the factor loading matrix to this triangular form, a unique global mode of the likelihood underlying the posterior distribution is guaranteed (Aßmann et al., 2016). However, it does not prevent the existence of local modes. Moreover, the constrains affect the shape of the likelihood and, consequently, the shape of the posterior distribution.

The pioneers of this method Anderson and Rubin (1956) viewed the rows of Λ as vectors in q space. The first row corresponds to the first coordinate axis. The second row lies in the plane regulated by the first two coordinate axis. The same rationale applies to the rest of the rows. Furthermore, they required that the matrix $\Lambda^T\Lambda$ be diagonal and its diagonal objects be positioned in decreasing order. The reason for this requirement derives from

the matrix theory and the following theorem. For a given positive definite matrix Λ, a uniquely determined orthogonal matrix Θ exists so that $\Theta^T\Lambda\Theta$ can be diagonal with diagonal elements arranged in decreasing order. If the matrix Λ is diagonal, then the uniquely determined orthogonal matrix Θ is the identity matrix I. Nevertheless, as Anderson and Rubin (1956) mention those conditions are not based on some theoretical support and they are somehow random. However, researchers ascertained that by implementing those restrictions, only local identification can be achieved leading to every factor loading matrix Λ having the same likelihood value with every factor loading matrix Λ' being a reflection of a subset of columns of the matrix Λ. This is known as transparent multimodality (Jennrich, 1978). In order to achieve identification and to deal with transparent multimodality researchers added the restriction that the diagonal elements of the factor loading matrix should be strictly positive.

The Lopes and West (2004) prior specification used for the Positive Lower Triangular method is the following

For the non-diagonal elements of the factor loadings matrix

$$\Lambda_{ij} \sim N(0, C_0)\ \forall i \neq j,\ i = 1, \ldots, p, j = 1, \ldots, q.$$

For the diagonal elements of the factor loadings matrix

$$\Lambda_{ii} \sim N(0, C_0)1(\Lambda_{ii} > 0),\ i = 1, \ldots, q.$$

This prior specification is a truncation of the normal distribution restricting the distribution to only positive values.

For the variances of the idiosyncratic errors

$$\sigma_i^2 \sim IG\left(\frac{v}{2}, \frac{vs^2}{2}\right),\ i = 1, \ldots, p.$$

The fully conditional posterior of the factors F is

$$F_i \sim N((I + \Lambda^T\Sigma^{-1}\Lambda)^{-1}\Lambda^T\Sigma^{-1}Y_i, (I + \Lambda^T\Sigma^{-1}\Lambda)^{-1}),\ i = 1, \ldots, n.$$

The fully conditional posterior of the elements of the variance of the idiosyncratic errors

$$\sigma_i^2 \sim IG\left(\frac{v+n}{2}, \frac{vs^2 + d_i}{2}\right)$$

$$\text{where } d_i = (Y_i - F\Lambda_i^T)^T(Y_i - F\Lambda_i^T),\ i = 1, \ldots, p.$$

The fully conditional for the rows $i = 1, \ldots, q$ of the factor loadings matrix

$$\Lambda_i \sim N(m_i, C_i)1(\Lambda_{ii} > 0)$$
$$\text{where } m_i = C_i(C_0^{-1}\mu_0 1_i + \sigma_i^{-2}F_i^T Y_i)$$
$$\text{and } C_i^{-1} = C_0^{-1}I_i + \sigma_i^{-2}F_i^T F_i.$$

The fully conditional for the rows $i = q+1, \ldots, p$ of the factor loadings matrix

$$\Lambda_i \sim N(m_i, C_i)$$
$$\text{where } m_i = C_i(C_0^{-1}\mu_0 1_q + \sigma_i^{-2}F_i^T Y_i)$$
$$\text{and } C_i^{-1} = C_0^{-1}I_q + \sigma_i^{-2}F_i^T F_i.$$

Finally, in order for Gibbs sampler to be implemented with the previously mentioned prior specification, the following steps must be followed

Step 1 Give initial values for $(\Lambda^{(0)}, \Sigma^{(0)}, F^{(0)})$.

Step 2 In each iteration $t = 1, \ldots T$

Step 2-a Update the rows $i = 1, \ldots, q$ of the factor loadings matrix

$$\Lambda_i^{(t)} \sim N(m_i, C_i)1(\Lambda_{ii}^{(t)} > 0)$$
$$\text{where } m_i = C_i(C_0^{-1}\mu_0 1_i + \sigma_i^{-2^{(t-1)}}F_i^{T^{(t-1)}} Y_i)$$
$$\text{and } C_i^{-1} = C_0^{-1}I_i + \sigma_i^{-2^{(t-1)}}F_i^{T^{(t-1)}} F_i^{(t-1)}.$$

Step 2-b Update the rows $i = q+1, \ldots, p$ of the factor loadings matrix

$$\Lambda_i^{(t)} \sim N(m_i, C_i)$$
$$\text{where } m_i = C_i(C_0^{-1}\mu_0 1_q + \sigma_i^{-2^{(t-1)}}F_i^{T^{(t-1)}} Y_i)$$
$$\text{and } C_i^{-1} = C_0^{-1}I_q + \sigma_i^{-2^{(t-1)}}F_i^{T^{(t-1)}} F_i^{(t-1)}.$$

Step 2-c Update the elements of the variance of the idiosyncratic errors

$$\sigma_i^{2^{(t)}} \sim IG(\frac{v+n}{2}, \frac{vs^2 + d_i}{2})$$
$$\text{where } d_i = (Y_i - F^{(t-1)}\Lambda_i^{T^{(t)}})^T(Y_i - F^{(t-1)}\Lambda_i^{T^{(t)}}),\ i = 1, \ldots, p.$$

Step 2-d Update the factors

$$F_i^{(t)} \sim N((I+\Lambda^{T^{(t)}}\Sigma^{-1^{(t)}}\Lambda^{(t)})^{-1}\Lambda^{T^{(t)}}\Sigma^{-1^{(t)}}Y_i, (I+\Lambda^{T^{(t)}}\Sigma^{-1^{(t)}}\Lambda^{(t)})^{-1}),\ i = 1, \ldots, n.$$

Finally, an important note regarding the later prior specification is that if we set the hyper parameter of the variance of the prior of the factor loading matrix to 0 and both the hyper parameters of the inverse Gamma prior distribution of the idiosyncratic errors to 0.001 then the prior specification coincides with the prior specification in the MCMCpack ("MCMCpack", n.d.). By setting the hyper parameter of the variance of the prior of the factor loading matrix to 0 we are employing a improper prior. Therefore, the Gibbs sampler of the MCMCpack of R (Team, 2021) is similar with the one described above. The same applies to the posterior distributions of the parameters of interest.

As far as the generation of the MCMC samples regarding the PLT method is concerned, in this thesis the MCMCpack ("MCMCpack", n.d.) will be used with the above-mentioned restrictions.

2.2 Bayesian Exploratory Factor Analysis

Bayesian Exploratory Factor Analysis (BEFA) is a pioneering method because it makes inference in models with simple dedicated structure and with correlated factors. Conti et al. (2014) define Dedicated Structure as the Perfect Simple Structure as described by Thurstone (1935). Dedicated Structure is the structure of a factor model where all measurements load onto one factor at most. Moreover, BEFA is an integrated Bayesian approach because it produces the posterior distribution of the parameters of interest, chooses the number of factors in the factor model and allocates the observed measurements to factors and to the equivalent factor loadings at the same time. Furthermore, in order to achieve identification, this method applies classical identifying criteria to the prior distribution of the parameters of the factor model. For identification to be achieved, the method restricts the sampler to stay in regions where only identified models are generated. The proposed realizations by this method have the same likelihood and they are similar up to column and sign switching. Consequently, BEFA solves the rotation problem and the uniqueness problem simultaneously. Furthermore, BEFA employs all of the accessible knowledge from the data by removing only measurements

which do not load on any factor. An important note about the method is that it can be used for continuous and binary observed measurements. BEFA has a dedicated method of generating MCMC samples which is highly modifiable according to the needs of the researcher. Before proceeding with the prior specification of the parameters and the description of the BEFA algorithm, some concepts have to be defined.

Conti et al. (2014) define the dedicated factor model with continuous and binary measurements the following way: Assuming that the p binary and continuous variables are arranged in a vector $(Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ for each observation $i = 1, 2, \ldots, n$. In order for Conti et al. (2014) to use one notation for the two data types, it is considered that every measurement is documented by a continuous latent variable $Y_{ip}^{\star}$.

$$Y_{ij} = \begin{cases} Y_{ij}^{\star} & \text{If } Y_{ij} \text{ is continuous} \\ 1[Y_{ij}^{\star} > 0] & \text{if } Y_{ij} \text{ is binary} \end{cases}, \; j = 1, \ldots p$$

The occurring continuous latent variable can be written as a linear combination of the k observed variables and q latent factors in the following form

$$Y_i^{\star} = BX_i + \Lambda F_i + \varepsilon_i$$

where B denotes the matrix of regression coefficients representing the effect of the covariates $X = (X_i, X_2, \ldots, X_n)$ on the latent variables $Y^{\star}$, Λ the factor loadings matrix, $F = (F_1, F_2, \ldots, F_n)$ the unobserved latent factors which follow a normal distribution with mean equal to 0, correlation matrix equal to R and covariance matrix of rank q equal to Z with diagonal elements equal to 1(the variance of each factors is equal to 1), $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ the idiosyncratic errors which follow a normal distribution with mean equal to 0 and variance equal to $\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$, where Σ is a diagonal matrix with its objects strictly positive, on the occasion that the latent variable $Y_i^{\star}$ is produced from a binary variable, then the variance of the specific idiosyncratic errors will be set equal to 1. Additionally, the idiosyncratic errors are mutually uncorrelated and the latent factors, the regression coefficients and the idiosyncratic errors are all independent from each other. Furthermore, Conti et al. (2014) assume that the factor model has a dedicated structure. An important note about the model is that on the condition that all the variables are continuous and the model does not include regression coefficients, then the dedicated factor model with continuous and binary measurements as defined by Conti et al. (2014) coincides with the typical basic q-factor

model (1.1).

Conti et al. (2014) denote as Δ the Binary indicator matrix with dimensions equal to that of the factor loadings matrix Λ and with the attribute to assign uniquely observed measurements to latent factors. Consequently, the rows of Δ denote that the equivalent observed measurement loads on the analogous latent factor. Before proceeding, it is important to illustrate an example with the binary indicator matrix in order to make it more perceptible.

For example, consider that the m-th row of the binary indicator matrix has the following form

$$\Delta_m = (0, 0, \ldots, 0, 0, \underbrace{1}_{\text{l-th element}}, 0, 0, \ldots, 0, 0).$$

This m-th row of the binary indicator matrix denotes that the m-th observed measurement is related to the l-th latent factor. In this occasion the m-th row of the binary indicator matrix is the indicator vector e_l. Because of the dedicated structure of the model, the measurements can load onto one factor utmost. The indicator vector of the rows of the measurements which does not load onto any factor(have a row full of 0) is denoted as e_0 and those measurements are dropped out of the model.

The binary indicator matrix plays a crucial role to the method by dictating which measurements are assigned to which factors, the number of factors with at least three measurements loaded onto them and which factors have less than three measurement loaded onto them and have to be removed from the model. The restriction of at least three measurements per factor originates from the need of the model to solve the uniqueness problem, which according to Anderson and Rubin (1956) is resolved when the number of factors is less than the Ledermann bound and each factor has at least three measurements loaded onto them. Finally, the binary indicator matrix is unknown and need to be estimated by the data.

As we have already mentioned, the generated samples of this method do not have uniqueness nor rotational issues. This statement can be supported by the following theorem suggested by Conti et al. (2014).

Assuming a dedicated factor model with q factors and covariance matrix of the factors of rank q and its diagonal elements equal to 1. Additionally, consider that in the l-th column of the binary indicator matrix Δ the number of

1 is either 0 or at least 3.

$$n_l(\Delta) = \sum_{m=1}^{p} \Delta_{ml} \begin{cases} n_l \geq 3 \\ n_l = 0 \end{cases} \quad \forall l = 1, \ldots, q.$$

Consequently, the factor model is identifiable up to sign and column switching. Those sign and column switching must be applied to the factor loadings matrix Λ and to the objects of the correlation matrix Z corresponding to columns of the binary indicator matrix Δ with at least 3 measurements. Finally, the same column switching must also be applied to the indicator matrix Δ. The objects on the correlation matrix Z corresponding to columns of the binary indicator matrix Δ with less than 3 measurements are not identified. Nevertheless, the identification of the covariance between the unidentified factors and the covariance between unidentified factors and identified factors are not important, because only the identified factors contribute information to the model.

Prior Specification

Conti et al. (2014) constructed the prior of the binary indicator matrix Δ having in mind the following philosophy. Conti et al. (2014) viewed the assignment of measurements to dedicated factors as a finite mixture with unknown number of components. Assume that t_l is the probability of an observed measurement to load on factor l and t_0 is the probability of an observed measurement not loading onto any factor. The previous sentence can be written in mathematical notation and it will have the following form

$$Pr(\Delta_m = e_l | t_l) = t_l \ , \ \sum_{l=0}^{q} t_l = 1 \ , \ l = 1, 2, \ldots, q$$

where Δ_m , for $m = 1, 2, \ldots p$ is the m-th row of the binary indicator matrix Δ and e_l is the indicator vector.

Conti et al. (2014) viewed the assignment of every measurements to dedicated factors as a two-stage procedure and employed a hierarchical prior on the binary indicator matrix Δ. At the first stage the measurement may be uncorrelated with the other measurements and not load to any factor. The probability of this is equal to t_0. Those measurements are removed from the model. Consequently, the probability of one measurement to load onto

one factor is equal to $1 - t_0$. In the second stage, conditionally on the measurement loading onto one factor, the probability of one measurement to be assigned to the l-th dedicated factor is equal to $t_l^\star$ for $l = 1, 2, \ldots, q$ where $\sum_{l=1}^{q} t_l = 1$. The probabilities of the different events can be expressed as

$$t = (t_0, t_1, t_2, \ldots, t_q) = (t_0, (1 - t_0)t_1^\star, (1 - t_0)t_2^\star, \ldots, (1 - t_0)t_q^\star)$$

with prior distributions for $t_0, t^\star$

$$t_0 \sim Beta(k_0, \xi_0) \text{ and } t^\star = (t_1^\star, t_2^\star, \ldots, t_q^\star) \sim Dir(k_1, k_2, \ldots, k_q).$$

However, the prior does not assure identification so far because it does not incorporate the restriction that each factor must have at least three dedicated measurements loaded onto them. In order for the Conti et al. (2014) to accomplish that, the distribution of Δ is limited to the sub-set of matrices D, belonging only to identifiable models.

$$P(\Delta|t, D) \propto \left(\prod_{l=0}^{q} t_l^{n_l(\Delta)} \right) \delta_D(\Delta)$$

where $n_l(\Delta) = \sum_{m=1}^{p} \Delta_{mq}$ denotes how many measurements are dedicated to factor l, for $l = 1, 2, \ldots, q$, $n_0(\Delta)$ denotes how many measurements do not load on any factor and $\delta_D(\Delta)$ denotes the Dirac measure which is a binary variable, which is equal to 1 on the occasion that Δ belongs to D and equal to 0 on the other occasion.

Conti et al. (2014) proposed a second more flexible hierarchical prior for the binary indicator matrix Δ which incorporates measurement-specific inclusion probabilities. In this prior specification each measurement has its own probability of inclusion t_{0m} for $m = 1, 2, \ldots, p$ which is independent from the other probability of inclusion. However, the probability of assigning one measurement to the l factor conditionally on the measurement loading onto one factor is the same as before. Conclusively, in this prior specification the probabilities of the different events can be expressed as

$$t^\star \sim Dir(k_1, k_2, \ldots, k_q) \ , \ t_m = (t_{0m}, (1 - t_{0m})t_1^\star, (1 - t_{0m})t_2^\star, \ldots, (1 - t_{0m})t_q^\star)$$

$$\text{for } m = 1, 2, \ldots, p \text{ and } t_{0m} \sim Beta(k_0, \xi_0).$$

Conti et al. (2014) proved in their simulated experiments that the second prior specification operates more efficiently in comparison to the first prior specification, because it is able to identify quite easily the uncorrelated measurements not loading onto any factor. Specifically, this can happen to a greater extent when the number of measurements is large.

For the prior distribution of the variance of the idiosyncratic error of the continuous measurements Conti et al. (2014) selected the Inverse-Gamma prior. The prior specification is the following

$$\sigma_m^2 \sim IG\left(c_0, C_m^0\right) \;\; m = 1, 2, \ldots, p.$$

Before proceeding with the specification of the prior distribution of the factor loadings it is important to clarify the impact of the binary indicator matrix to the factor loadings matrix. To demonstrate that, consider a factor loading Λ_{ml} in row m and column l. This factor loading will be either equal to 0 on the occasion that the object of the binary indicator matrix Δ on the m row and l column is equal to 0, or follow a prior distribution when the object of the binary indicator matrix Δ on the m row and l column is equal to 1. Conti et al. (2014) assumed that the factor loadings are independent across measurements. They selected as prior distribution for the factor loadings the normal distribution on condition that the idiosyncratic variance is known and the object in the binary indicator matrix being in the same row and the same column as the factor loading is equal to 1. Conclusively, the factor loadings matrix Λ will have in the m-th row a single non-zero factor loading Λ_m^{Δ} which follows conditionally the normal distribution

$$\Lambda_{ml}^{\Delta} | \sigma_m^2 \sim N(a_m^0, A_m^0 \sigma_m^2)$$

where a_m^0 denotes the prior mean and A_m^0 the scale of the variance.

For each of the rows of the matrix of regression coefficients Conti et al. (2014) selected as prior distribution the normal distribution

$$B_m \sim N(b_m^0, M_m^0) \text{ for } m = 1, 2, \ldots, p$$

where b_m^0 is the prior mean and M_m^0 is the prior covariance matrix.

Data augmentation is employed by Conti et al. (2014) in order to generate samples of the correlation matrix of factors. Before proceeding with the description of this procedure, it is crucial to illustrate the link between the distribution of the covariance matrix of factors Z, the distribution of the correlation matrix of factors R and the distribution of the variance of the factors a.

The covariance matrix Z can be decomposed as $Z = a^{\frac{1}{2}} R a^{\frac{1}{2}}$, where $a = diag(a_1, a_2, \ldots, a_q)$ denotes a diagonal matrix with elements the variances of the factors and R denotes the correlation matrix of the factors. Zhang, Boscardin, and Belin (2006) proved that on the occasion that the covariance matrix of factor Z follows an Inverse-Wishart distribution $Z \sim IW(v, S)$, where v denotes the degrees of freedom having the following constraint $v - q + 1 > 0$ and S denotes a scale matrix. Then by employing the Jacobian transformation from $Z \to (a, R)$ the joint distribution of (a, R) can be acquired. The joint distribution of (a, R) has the following form

$$P(a, R|S) = J(Z \to (a, R))P(Z) = c|S|^{\frac{v}{2}}|a|^{-\frac{v}{2}-1}|R|^{\frac{-(v+q+1)}{2}} \exp\left[-\frac{1}{2} trace(Sa^{-1}R^{-1})\right].$$

An important note regarding the scale matrix hyper parameter S of the inverse-Wishart prior is that it can be selected to have either fixed values or follow a hyper prior distribution $P(S)$. Conti et al. (2014) suggested that the latter option should be optimal and selected as hyper prior of the scale matrix hyper parameter S the prior created by Huang and Wand (2013). The selection of S as random(follow a hyper prior) is made because when S is random, the marginal data augmentation algorithm operates move efficiently and has better mixing. According to the prior proposed by Huang and Wand (2013), the scale matrix $S = diag(S_1, S_2, \ldots, S_q)$, is a non-singular diagonal matrix with each element S_j for $j = 1, 2, \ldots, q$ following a gamma distribution

$$S_l \sim G\left(\frac{1}{2}, \frac{1}{2v^\star, A_l^2}\right) \text{ for } l = 1, 2, \ldots, q \text{ where } v^\star = v - q + 1.$$

According to Zhang et al. (2006) the marginal distribution of R can be calculated in closed form by integrating out a, when the scale matrix $S =$

$diag(S_1, S_2, \ldots, S_q)$ is a non-singular diagonal matrix regardless of S being fixed or either following a hyper prior (being random). Consequently, the prior distribution of R is not affected by S whether fixed or following a hyper prior. The marginal distribution of R by integrating out a has the following form

$$P(R|S) = \int P(a, R|S)da = 2^{v\frac{q}{2}}\Gamma(\frac{v}{2})|R|^{-\frac{(v+l+1)}{2}}(\prod_l r^{ll}) = P(R).$$

An important note about the marginal density $P(R)$ of the correlation matrix of factors R is that it is independent from S making the degrees of freedom v as the only hyper parameter of marginal density. Barnard, McCulloch, and Meng (2000) proposed how to specify the degrees of freedom v for the Inverse-Wishart distribution. Barnard et al. (2000) also proved that the selection of v equal to $v = q + 1$ makes the individual correlations follow marginally the uniform distribution. Furthermore, Barnard et al. (2000) proved that as the number of degrees of freedom v grow, the distribution of the individual correlations takes a ball-shape form and bounds the individual correlations away from neighbourhoods close to ± 1.

A major role in the tuning of the algorithm of the method proposed by Conti et al. (2014) is played by the degrees of freedom v of the Inverse-Wishart distribution. On the occasion that the prior distribution enables large correlation between the latent factors, then it is highly possible that the method may select a large number of factors which will be created by some factors split to many highly correlated factors.

Conclusively, In order for the marginal data augmentation algorithm to operate, it needs the generation of the variance of the factors a from the conditional distribution $P(a|R)$ for a pre-specified value of R. On the occasion that S follows a hyper prior, then the variance of the factors a is generated from the joint prior $P(a, S|R) = P(a|S, R)P(S|R) = P(a|S, R)P(S)$ where $P(S|R) = P(S)$ and $P(S)$ is equivalent to the prior of S. The conditional distribution of $P(a|R, S)$ can be derived from $P(a|R, S)$ by doing the following calculations $P(a|R, S) = \frac{P(a,R|S)}{P(R|S)} = \frac{P(a,R|S)}{P(R)}$. It can be easily proven that every variable $a_l|S_l, R$ follows an Inverse-Gamma distribution with each S_l following a pre-specified hyper prior which in the method of Conti et al. (2014) is the prior proposed by Huang and Wand (2013).

$$S_l \sim G\left(\frac{1}{2}, \frac{1}{2v^\star, A_l^2}\right) \ , \ a_l|RS_l \sim IG\left(\frac{v}{2}, \frac{s_l r^{ll}}{2}\right).$$

On the occasion that S is pre-specified values, then $a_l|R, s_l$ is generated conditionally on that value.

How the MCMC sampling scheme of Bayesian Explanatory Factor Analysis operates

BEFA MCMC sampling scheme can be divided into two parts. The first part is the unrestricted MCMC sampler. In this part, the limitation of at least three measurements dedicated to each factor is not applied and the samples are generated conditionally only on the latest generated values of the other parameters and latent factors. This results in good mixing in the Markov Chain at the expense of the proposed samples not being identifiable. Those non identified samples are not used for posterior inference and their only purpose is to come upon models with dissimilar dimensions in order to create relevant candidate moves for the M-H moves of the second part of the sampling scheme. The algorithm of the unrestricted MCMC sampler is the following

A Generate the binary indicator matrix Δ, the factor loadings Λ and the idiosyncratic variances Σ at the same time.

A 1 Generate the binary indicator matrix Δ row-wise by employing Gibbs updates. The posterior probability of the m-th measurement being dedicated to the l-th factor is a function of the marginal likelihood of the equivalent latent variable, for $l = 0, 1, 2, \ldots, q$

$$Pr(\Delta_m = e_l|Y^\star_{\cdot m}, \Delta_{-m}, X, F, B_m, t) =$$
$$\frac{P(Y^\star_{\cdot m}|\Delta_m = e_l, X, F, B_m) * P(\Delta_m = e_l|\Delta_{-m}t)}{\sum_{k=0}^{q} P(Y^\star_{\cdot m}|\Delta_m = e_k, X, F, B_m) * P(\Delta_m = e_k|\Delta_{-m}t)}$$

where $P(Y^\star_{\cdot m}|\Delta_m = e_l, X, F, B_m)$ denotes the marginal likelihood of the vector $Y^\star_{\cdot,m} = (Y^\star_{1m}, Y^\star_{2m}, \ldots, Y^\star_{nm})$, conditioning on the rest of the rows Δ_{-m} of the binary indicator matrix Δ.

A 2 Generate at the same time for each measurement m, the idiosyncratic variance Σ on the occasion that the measurement is continuous and the factor loadings Λ. Denote as Λ^Δ_m the only non-zero element in row m of the

factor loadings matrix Δ.

On the occasion that a measurement is continuous and does not load on any factor in a dedicated way, generate the idiosyncratic variance as follows

$$\sigma_m^2|\cdots \sim IG(c_n, C_m^{nN})$$
$$c_n = c_0 + \frac{n}{2} \text{ and } C_m^{nN} = c_m^0 + \frac{\tilde{Y}_{\cdot m}^T \tilde{Y}_{\cdot m}}{2}$$
$$\tilde{Y}_{\cdot M} = Y_{\cdot m}^\star - XB_m.$$

On the occasion that a measurement loads on one factor in a dedicated way, generate the idiosyncratic variance and the non zero factor loading as follows

$$\sigma_m^2|\cdots \sim IG(c_n, C_m^n)$$
$$\Lambda_m^\Delta|\sigma_m^2, \cdots \sim N(A_m^n a_m^n, A_m^n \sigma_m^2)$$
$$c_n = c_0 + \frac{n}{2}$$
$$C_m^n = C_m^0 + \frac{1}{2}\left(\tilde{Y}_{\cdot m}^T \tilde{Y}_{\cdot m} + \frac{(a_m^0)^2}{A_m^0} - A_m^h (a_m^h)^2\right)$$
$$(A_m^n)^{-1} = (A_m^0)^{-1} + \sum_{i=1}^{n} F_{il}.$$

B Generate the regression coefficients B row-wise by employing the following conditional distribution

$$B_m \sim N(M_m^n b_m^n, M_m^n)$$
$$(M_m^h)^{-1} = (M_m^0) + \frac{1}{\sigma_m^2} X^T X$$
$$b_m^n = (M_m^0)^{-1} b_m^0 + \frac{1}{\sigma_m^2} X^T (Y_{\cdot m}^\star - F\Lambda_m)$$
$$m = 1, \ldots, p$$

where Λ_m denotes the m-row of the factors loadings matrix Λ.

C On the occasion that m is dichotomous, its latent variable $Y_{im}^{\star}$ is generated by employing the following distribution

$$Y_{im} \sim \begin{cases} TN_{(-\infty,0]}(X_i^T B_M + F_i^T \Lambda_m, 1) & \text{if } Y_{im} = 0 \\ TN_{(0,+\infty)}(X_i^T B_M + F_i^T \Lambda_m, 1) & \text{if } Y_{im} = 0 \end{cases}$$
$$i = 1, \ldots, n$$

where TN denotes a Truncated Normal distribution.

D Generate the factors F and their correlation matrix R jointly by employing Marginal Data Augmentation.

The method has a tuning parameter K which indicates the maximum number of factors, however, some of those factors may not have measurements loaded onto them. Conti et al. (2014) divide the factors into two categories. The first category is the active factors K_1 with at least three measurements loaded onto them. The second category is the inactive factors K_2 with less than three measurements loaded onto them. The active factors indicate the non-zero columns of the factors loadings matrix Λ and the inactive factors indicate the zero columns. At a given MCMC iteration the inactive factors can be turned to factors on the occasion that at least three measurements are loaded onto them. Likewise, the active factors can turn into inactive when less that 3 measurements are loaded onto them.

Suppose that in a iteration of the algorithm, there are K_1 active factors and K_2 inactive factors where $K = K_1 + K_2$. The method rearranges the factors in such a way that the active factors are in the first positions and the inactive factors in the last positions. Consequently, the rows and columns of the other parameters are rearranged in the following way

$$F = (F_1, F_2) \ , \ \Lambda = (\Lambda_1, \Lambda_2)$$

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \ , \ Z = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix}$$

where R denotes the correlation matrix of the factors and Z the covariance matrix of the factors. When the method generates the active and inactive factor and their correlation matrix at the same time, then it results in the generated correlation of the inactive factors having high autocorrelation between MCMC iterations and also poor mixing. Consequently, sampling in

this fashion does not enable the search procedure to select new active factors from the inactive ones, because all of which are almost identical.

In order to solve the previously mentioned problem, the method generates the inactive factors and the covariance matrix Z in the same time in the marginal data augmentation procedure. The sampling strategy can be divided into two parts. In the first part, because the likelihood is independent of the inactive factors, the active factors can be updated marginally by integrating out the inactive factors (Van Dyk & Park, 2008). F_{i1} marginal conditional prior distribution is $N(0, Z_{11})$, and for all $i = 1, \ldots, n$ the updated conditional posterior is calculated as

$$F_{1i}|Z_{1i} \sim N(A_{F_i}\Lambda_{F_{i1}}, AF_1)$$

$$\text{where } (A_{F_1})^{-1} = \Lambda_1^T(\Sigma)^{-1}\Lambda_1 + (Z_{11})^{-1} \text{ and } \Lambda_{F_1} = \Lambda_1^T(\Sigma)^{-1}(Y_i^\star - BX_i).$$

After the update of F_1, the inactive factors and the covariance matrix can be generated at the same time. In the expanded model, their joint distribution, is the following

$$P(F_2, Z|F_1, Y, \Lambda, \Sigma) \propto P(F_2|Z, F_1)P(Z_{12}, Z_{22}|Z_{11})P(Z_{11}|F_1).$$

The previous equation makes clear that the covariance matrix Z can be generated in blocks. Having this is mind, Conti et al. (2014) created a sampling scheme based on some properties of the Inverse-Wishart distribution. Specifically, the property of Inverse-Wishart distribution stating that the matrix Z_{11} is independent of the block matrices $Z_{11}^{-1}Z_{12}$ and $Z_{22.1}$. $Z_{22.1} = Z_{22} - Z_{21}Z_{11}^{-1} - Z_{12}$ denoting the Schur complement of Z_{11} in Z both a priori and a posteriori. Consequently, the scale matrix S in the inverse-Wishart prior and posterior distribution of Z can be divided in the same way as Z in the following form

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

The method employing the prior $Z_{11} \sim IW(v - K_2, S_{11})$, proceeds with the generation of the block matrix Z_{11} conditional on F_1 as follows

$$Z_{11}|F_1 \sim IW(v - K_2 + n, S_{11} + F_1^T F_1).$$

Afterwards, the Schur complement $Z_{22.1}$ and the product $Z_{11}^{-1}Z_{12}$ are generated jointly as follows

$$Z_{22.1} \sim IW(v, S_{22.1})$$
$$Z_{11}^T Z_{12}|Z_{22.1} \sim N(S_{11}^{-1}S_{12}, S_{11}^{-1} \otimes Z_{22.1}).$$

Finally, after the generation of the different blocks, the inactive factors are generated independently for $i = 1, 2, \ldots, n$ from the conditional distribution $P(F_2|Z, F_1)$ as follows

$$F_{21}|Z, F_{1i} \sim N((Z_{11}^{-1}Z_{12})^T F_{1i}, Z_{22.1}).$$

E Generate the indicator probabilities t by first generating the components t_{0m} and $t^\star$

$$t_{0m} \sim Beta(k_0 + 1[\Delta_m = e_0], \xi_0 + p - [\Delta_m = e_0])$$
$$t^\star \sim Dir(k_1 n_1(\Delta), k_2 n_2(\Delta), \ldots, k_q n_q(\Delta))$$

where $n_l(\Delta) = \sum_{m=1}^{p} 1[\Delta_m = e_l]$ denotes how many measurements loads only to factor l. Afterwards calculate the resulting probabilities t_m as

$$t_m = (t_{0m}, (1 - t_{0m})t_1^\star, (1 - t_{0m})t_2^\star, \ldots, (1 - t_{0m})t_q^\star).$$

The second part of the BEFA MCMC sampling scheme is the M-H moves with intermediate steps in non identified models. This step follows the philosophy of marginal data augmentation. Before proceeding with the explanation of the second part of the BEFA MCMC sampling scheme, it is important to highlight the similarities and differences between the marginal data augmentation and the M-H moves with intermediate steps in non identified models. In order to improve the sampling and mixing both methods make intermediate steps in non-identifiable models. However both methods secure the return to identifiable models after having completed those intermediate steps. Marginal data augmentation at each MCMC iteration expands the parameter space by adding a set of working parameters which are not part of the original model and most of the times, the identification from the data is not possible. After the introduction of the working parameters and the

transformation of the model a Gibbs sweep is executed. Afterwards, the model re-transforms to its original form. Upon every MCMC iteration, the M-H moves with intermediate steps in non identified models makes steps in possible non-identifiable models by employing the unrestricted MCMC sampler for a pre-specified(or stochastic) number of sweeps in order to generate a candidate update for the parameters. This candidate update is accepted as the new state of the model if the proposed binary indicator matrix belongs to the sup-set of matrices D. An important note in both methods is that the values of the parameters generated by the sweeps in the intermediate steps are not saved for inference. Only the values of the parameters when the model is re-transformed are saved for inference. Finally, the key difference between the two methods is that the marginal data augmentation introduces extra parameters(Working parameters) and M-H moves with intermediate steps in non identified models just loosens some restrictions. Specifically, the identification restriction requiring each factor to have at least three measurements loaded on them. The algorithm of M-H moves with intermediate steps in non identified models sampler is the following

Denoting as $\Theta = (Y^{\star}, F, \Delta, \Lambda, B, \Sigma, R, t)$ the set of model parameters to be generated. The method at each MCMC iteration executes a predetermined (or stochastic) number of sweeps with the unrestricted MCMC sampler which temporarily enables the Markov Chain to come across possible non identifiable states of the model in order to create a candidate which will be accepted (or rejected) as the new state of the algorithm by a M-H step. Assuming that the present state of the algorithm is $\hat{\Theta}_0$, a candidate state $\check{\Theta}_0$ is created by executing $2S$ intermediate MCMC sweeps of the unrestricted MCMC sampler as follows

Step 1 Having as starting state the state $\hat{\Theta}_0$ generate a sequence $(\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_{S-1}, \hat{\Theta}_S) = \bar{\Theta}_S$ by executing S sweeps of the unrestricted MCMC sampler in the ordinary order (A 1, A 2, A 3, B, C, D, E).

Step 2 Having as starting state the state $\check{\Theta}_S = \bar{\Theta}_S$ generate the sequence $(\check{\Theta}_{S-1}, \check{\Theta}_{S-1}, \ldots, \check{\Theta}_2, \check{\Theta}_1, \check{\Theta}_0)$ by executing S sweeps of the unrestricted MCMC sampler in reverse order (E, D, C, B, A 1, A 2, A 3).

Step 3 The candidate state $\check{\Theta}_0$ is adopted as the new state on the occasion that its candidate binary indicator matrix $\check{\Delta}_0$ belongs to the sup-set of matrices D. Otherwise, the old state of the chain $\hat{\Theta}_0$ becomes the new state.

An important note about the method is that the need for the computation of

the normalizing constant in $P(\Delta|t, D) \propto \prod_{l=0}^{q} t_l^{n_l(\Delta)} \delta_D(\Delta)$ becomes unnecessary because of the symmetrical intermediate steps. To be more precise, a candidate state $\check{\Theta}_0$ becomes the new state of the model if the candidate binary indicator matrix $\check{\Delta}_0$ belongs to the sup-set of matrices D. Otherwise, the candidate state is not accepted as the new state. Another important note is that the number $2S$ of intermediate steps controls the time the algorithm spends exploring the different models in order to create proposals for the M-H moves. Its tuning is of great importance for the convergence of the algorithm. Finally, it can have a predetermined value or change stochastically.

Bayesian Exploratory Factor Analysis method of solving the label switching problem

As already mentioned the method produces realizations which have the same likelihood and are identical up to sign and column switching. Nevertheless, those problems must be solved in order to make inference on the factor loadings matrix. Conti et al. (2014) proposed the following strategy in order to solve the label switching problem. First, Conti et al. (2014) solve the column switching problem by rearranging the columns according to the top element of the columns. Every top element in each active column corresponds to a different measurement on account of the dedicated structure of the factor loadings matrix. The non-zero columns of Λ are rearranged in such a way so that the top object appears in ascending order($l_1 < l_2 < \cdots < l_q$ where l_k denotes the first row in each active column k) in each MCMC iteration. Furthermore, appropriate rearrangements must be applied to the columns and rows of the correlation matrix of the factors R. Afterwards, to solve the sign switching problem Conti et al. (2014) employ a point of reference factor loading in each column,(the factor loading having the highest posterior probability of being non-zero in each column). Subsequently, the signs are changed in every iteration so that it can match with those of the point of reference.

2.3 Weighted Orthogonal Procrustes

Weighted Orthogonal Procrustes (WOP) was created by Aßmann et al. (2016) and is a post-hoc method created to solve the identification issue regarding the factor loadings matrix. The Weighted Orthogonal Procrustes (WOP) and Orthogonal Procrustes (OP) methods can be employed with every way

of generating MCMC samples of the parameters of the factor model, so researchers can use any method they want to generate samples as long as the generated factor loadings matrix is unconstrained. In order to generate samples of the parameters of the factor analysis model, WinBUGS(Lunn et al., n.d.) or OpenBUGS(Surhone et al., 2010) software can be used. Another option would be the programming language R (Team, 2021) using the packages Stan (Carpenter et al., 2017) or MCMCpack ("MCMCpack", n.d.). For the generation of the MCMC samples in this thesis regarding the WOP and OP methods, the MCMCpack will be used.

The Weighted Orthogonal Procrustes method was created in order to solve the identification issues in dynamic and static factor models. In order to do so, the method takes a Decision Theoretic approach where it is required to minimize a posterior expected loss. To minimize the posterior expected loss the method transforms the unconstrained Gibbs samples using a series of orthogonal matrices which is the analogue of using orthogonal Procrustes transformation on a factor model with no vector autoregressive dynamics. Before proceeding with a detailed explanation of the method, some concepts have to be defined. A dynamic factor model with n variables, time dimension T and q factor should have the following form

$$Y_t = \Lambda_0 F_t + \Lambda_1 F_{t-1} + \cdots + \Lambda_S F_{t-S} + e_t \ , \ t = 1, \ldots, T$$

where the factors follow an autoregressive process of order K with the following form

$$F_t = \Phi_1 F_{t-1} + \Phi_2 F_{t-2} + \cdots + \Phi_K F_{t-K} + \varepsilon_t$$

where Y_t denotes the stationary observed data, F_t the latent factors, Λ_s the factor loadings matrix and e_t the errors which are independently and identically distributed following normal distribution with diagonal covariate matrix $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$ and mean equal to 0. Additionally, $\Phi_k \ k = 1, \ldots, K$ denotes the persistence matrices of dimension $q \times q$ and ε_t the errors which are independently and identically distributed and follow the normal distribution with covariate matrix equal to the identity matrix I and mean equal to 0. An important note about the dynamic factor model is that for $S = 0$ and $K = 0$ the dynamic factor model coincides with the static factor model.

The identification issue regarding the matrix of factor loadings is resolved when it is certain that the parameters of interest obtained from the posterior

distribution are unique. As it has already been mentioned, in order to solve the issue, the WOP method takes a decision theoretic approach utilizing a loss function to evaluate the dissimilarity between the parameters Θ and their estimators $\Theta^{\star}$. A loss function $L(\Theta^{\star}, \Theta)$ defined by Jasra, Holmes, and Stephens (2005) is the mapping of the estimators $\Theta^{\star}$ belonging to the set of all feasible estimators Ξ and every one of the parameter values Θ in the parameter space on the line of the real numbers. With regard to the minimal expected loss, the optimal estimator is

$$\tilde{\Theta}^{\star} = arg \min_{\Theta^{\star}} \int_{\Theta} L(\Theta^{\star}, \Theta) P(\Theta|Y) d\Theta.$$

The weighted orthogonal Procrustes method adjusts the loss function in order to expand it and make it able to discriminate between invariant loss of estimators due to the post-multiplication of the factor loadings matrix with an orthogonal matrix D. The new loss function has to have the following form

$$L(\Theta^{\star}, \Theta) = \min_{D} \left[L_D(\Theta^{\star}, H(D)\Theta) \right] \ , \ D^T D = I$$

where $L_D(\Theta^{\star}, H(D)\Theta)$ is the loss due to post-multiplication of the factor loadings with an orthogonal matrix D for a given value of $\Theta^{\star}$. The loss function functional form is chosen to restrain the solutions of the particular minimization problem to the set of solutions which guarantee solvability and uniqueness. The loss function functional form is the following

$$L_D(\Theta^{\star}, H(D)\Theta) = \left(H(D)\Theta - \Theta^{\star}\right)^T \left(H(D)\Theta - \Theta^{\star}\right).$$

The integral in the expected posterior loss is calculated using Monte Carlo simulation methods. Therefore, the minimization problem can be rewritten as

$$\left[\left[\tilde{D}^{(r)} \right]_{r=1}^{R}, \tilde{\Theta}^{\star} \right] = argmin_{\left[\tilde{D}^{(r)}\right]_{r=1}^{R}, \Theta^{\star}} \sum_{r=1}^{R} L_D(\Theta^{\star}, H(D^{(r)})\Theta^{(R)}) \ , \ D^{(r)T} D^{(r)} = I \ , \ r = 1, \dots R$$

where $\Theta^{(r)}$, , $r = 1, \dots, R$ is the unconstrained posterior distribution. The key idea of the weighted orthogonal Procrustes method in order to solve the minimization problem is the following. The unconstrained sampler generates a sample from the posterior distribution $\left[\Theta^{(r)}\right]_{r=1}^{R}$ which can be rewritten as a

posterior distribution sample transformed by a random sequence of orthogonal matrices $\left[H(D^{(r)})\Theta^{(r)}\right]_{r=1}^{R}$. All the samples of the form $\left[H(D^{(r)})\Theta^{(r)}\right]_{r=1}^{R}$ have the same posterior probability. Every different sequence of orthogonal matrices $\left[D^{(r)}\right]_{r=1}^{R}$ leads to a different estimate of Θ. Because the expected loss function is globally convex in $\Theta^{\star}$, the all minima can be calculated as

$$\frac{1}{R}\sum_{r=1}^{R}\left(H(D^{(r)})\Theta^{(r)}-\overline{H(D)\Theta}\right)^{T}\left(H(D^{(r)})\Theta^{(r)}-\overline{H(D)\Theta}\right)$$

$$\text{where } \overline{H(D)\Theta}=\frac{1}{R}\sum_{r=1}^{R}H(D^{(r)})\Theta^{(r)}.$$

Consequently, for the $\left[H(D^{(r)})\Theta^{(r)}\right]_{r=1}^{R}$ samples the identification issue is resolved.

A solution for the identification issue regarding dynamic factor models with $S \geq 0$ and $K = 0$

In order to solve the optimization problem in a dynamic factor model with $S \geq 0$ and $K = 0$, the following iterative two-steps optimization procedure should be followed(The solution is explained in grater detail in Aßmann et al. (2016))

Initialization For starting values of $\Theta^{\star}$, choose the last generated sample from the unconstrained sampler.

I For each $r = 1 \ldots, R$ and for specified $\Theta^{\star}$ solve the following minimization problem

$$D^{(r)} = arg\min_{D} L_D(\Theta^{\star}, H(D)\Theta^{(r)})$$

$$\text{where } L_D(\Theta^{\star}, H(D)\Theta^{(r)}) = trace\left[\left(\bar{\Lambda}^{(r)}D-\bar{\Lambda}^{\star}\right)^{T}\left(\bar{\Lambda}^{(r)}D-\bar{\Lambda}^{\star}\right)\right]$$

where $\bar{\Lambda}^{\star}$ is the estimator of the stacked matrix of loadings $\bar{\Lambda} = (\Lambda_0^T, \ldots, \Lambda_S^T)^T$ and $\bar{\Lambda}^{(r)}$ is a generation of $\bar{\Lambda}$ created from the the unconstrained sampler.

This minimization problem can be solved by executing the following steps

I Step A Define $S_r = \bar{\Lambda}^{(r)T} W \bar{\Lambda}^{\star}$

where W denotes the weighting matrix which is a diagonal matrix with diagonal elements strictly positive. So the weights are a function of the size of the factors (how many factor we have) and the determinants of the estimated covariance matrices. The initial values of the weighting matrix W is equal to the $\frac{1}{l(q)}$ where $l(q)$ is the length of the factor loadings matrix. An important note about the weighting matrix is that if we set it equal to the identity matrix, then we have the orthogonal procrustes method.

$$W = R(\sum_{r=1}^{R} \sqrt{\bar{\Lambda}^{(r)} \bar{\Lambda}^{(r)T} \odot I})^{-1}$$

$$W = diag(w_1, \ldots, w_{(S+1)n})$$

$$\text{where } w_i = det\left(\frac{1}{R}\sum_{r=1}^{R}\left(\bar{\lambda}_i^{(r)} - \bar{\lambda}_i^{\star}\right)^T\left(\bar{\lambda}_i^{(r)} - \bar{\lambda}_i^{\star}\right)\right)^{-\frac{1}{q}} \quad i = 1, \ldots, (S+1)n$$

where $\odot$ is the Hadamard or Schur product as defined in Lutkepohl (JR, 1998), $\bar{\lambda_i}^{(r)}$ is the i-th row of $\bar{\Lambda}^{(r)}$ and $\bar{\lambda_i}^{\star}$ is the i-th row of $\bar{\Lambda}^{\star}$

I Step B Calculate the singular value decomposition $S_r = U_r M_r V_r^T$ where U_r is the matrix of eigenvectors of eigenvectors of $S_r S_r^T$, V_r is the matrix of eigenvectors of $S_r^T S_r$ and M_r is a diagonal matrix with elements the square roots of the eigenvalues of $S_r S_r^T$ and $S_r^T S_r$ which is the same.

I Step C Acquire the orthogonal matrix $D^{(r)} = U_r V_r^T$.

II Select $\bar{\Lambda}^{\star}$ and $\Sigma^{\star}$ as suggested by $\overline{H(D)\Theta}$.

III Terminate, when the sum of squared deviations between two consecutive $\Theta^{\star}$ is less than a predetermined threshold(Aßmann et al. (2016) suggested as threshold the value 10^{-9}).

2.4 Rotation Sign Permutation

Rotation Sign Permutation post processing algorithms (RSP) were created by Papastamoulis and Ntzoufras (2022a) and are post-hoc methods created to solve the identification issue regarding the factor loadings matrix. The

RSP methods can be used with every method generating MCMC samples corresponding to the parameters of the factor analysis model. In this way, any method can be used by researchers according to their needs so that samples can be generated provided that the generated factor loadings matrix is unconstrained. In order to generate samples of the parameters of the factor analysis model, WinBUGS(Lunn et al., n.d.) or OpenBUGS(Surhone et al., 2010) software can be used. Another option would be the programming language R (Team, 2021) using the packages Stan (Carpenter et al., 2017) or MCMCpack ("MCMCpack", n.d.). For the generation of the MCMC samples in this thesis regarding the RSP methods the MCMCpack ("MCMCpack", n.d.) will be used.

Before explaining in greater detail the RSP methods, some concepts have to be define. Lets define as T_q the set with all $q!$ permutations, as $v = (v_1, v_2, \ldots, v_q) \in T_q$ a permutation, as P the permutation matrix which has one value of 1 in each row and each column and everywhere else it has 0 and has dimensions $q \times q$. One feature of the permutation matrix is that it illustrates a permutation of the q elements and when it is multiplied with another matrix, it changes its row(pre-multiplied) or columns(post-multiplied). Another feature of the permutation matrix is that it is a rotation matrix. Lets also define as column-permuted matrix $\dot{\Lambda} = \Lambda P$ a loadings matrix post-multiplied with a permutation matrix. Before proceeding with the definition of the concepts, it is important to illustrate an example with the permutation matrix in order to make it more perceptible.

For example, lets assume a factor loadings matrix Λ with 3 factors(columns) and a permutation $v = (2, 3, 1) \in T_3$. This permutation coincides with the following permutation matrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Therefore, a column-permuted matrix with the previous permutation will have the following form

$$\dot{\Lambda} = \Lambda P = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \vdots & \vdots & \vdots \\ \lambda_{p1} & p_{32} & p_{33} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \lambda_{13} & \lambda_{11} & \lambda_{12} \\ \vdots & \vdots & \vdots \\ \lambda_{p3} & \lambda_{p1} & \lambda_{p2} \end{pmatrix}.$$

Lets define as signed permutation matrix Q a square matrix having one non-zero element in each row and column. This non zero element is either 1 or -1. A $q \times q$ signed permutation matrix is written as

$$Q = SP$$

where S is a $q \times q$ diagonal matrix in which each diagonal element is either 1 or -1. And P is a $q \times q$ permutation matrix. Before proceeding it is important to illustrate an example with the sign permutation matrix in order to make it more perceptible.

For example, lets assume that the factor loading matrix and permutation are the same as the previous example. Consequently, the permutation matrix will be the same as the previous example as well since permutation remains the same. Moreover, assuming that the diagonal matrix is the $S = diag(-1, -1, 1)$, then the sign permutation matrix of this example will have the following form

$$Q = SP = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix}.$$

And the factor loading matrix post-multiplied with the sign permutation matrix will have the following form

$$\Lambda^{\circ} = \Lambda Q = \Lambda SP = \begin{pmatrix} \lambda_{13} & -\lambda_{11} & -\lambda_{12} \\ \vdots & \vdots & \vdots \\ \lambda_{p3} & -\lambda_{p1} & -\lambda_{p2} \end{pmatrix}.$$

For the rest of this section, the following notation will be used in order to refer to the different forms of the factor loading matrix: $\tilde{\Lambda}$ Varimax rotated factor loading matrix, $\dot{\Lambda}$ column permuted factor loading matrix, Λ° sign and column permuted factor loading matrix and as $\Lambda^{\star}$ the reference matrix

of factor loadings.

The RSP methods philosophy of solving the identification issues of the factor loading matrix is the following: for each MCMC iterations the methods use the Varimax rotation in order to achieve a simple structure. However, the MCMC samples are still non identifiable due to sign and column switching. In order to solve this, the methods use the following strategy: signed permutation is applied to the MCMC samples until the transformed factor loading matrices are close to the reference matrix of factor loadings $\Lambda^\star$. If we define as Q_q the set of all $q \times q$ discrete signed permutation matrices, the optimization problem of solving the non identifiable due to sign and column switching can be summarized as

$$\begin{aligned} &\text{Minimize } \sum_{t=1}^{T} \left\| \tilde{\Lambda}^{(T)} Q^{(T)} - \Lambda^\star \right\|^2 \\ &\text{Subject to } Q^{(t)} \in Q_q \; t = 1, \dots T \end{aligned}$$

where T is the length of the MCMC chain after the burn in and $\|A\| = \sqrt{\sum_i \sum_j a_{ij}^2}$ is the Frobenius norm on the matrix space. The above optimization problem subject to $Q^T Q = I_q$ is known as the Orthogonal Procrustes problem, a solution achieved using the singular value decomposition $\Lambda^\star(\Lambda^{(t)})^T$ (Schönemann, 1966). The above Frobenius norm in the minimization can be expressed as

$$\left\| \tilde{\Lambda} Q - \Lambda^\star \right\|^2 = \left\| \tilde{\Lambda} S P - \Lambda^\star \right\|^2 = \sum_{r=1}^{p} \sum_{j=1}^{q} \left(s_j \tilde{\lambda}_{r v_j} - \lambda^\star_{rj} \right).$$

If we denote as $v = (v_1, \dots, v_q) \in T_q$ the permutation vector analogous to the permutation matrix P, then the minimization can be written as

$$min \Psi(\Lambda^\star, s, v) = \sum_{t=1}^{T} L^{(t)}_{s^{(t)}, v^{(t)}} \tag{2.1}$$

$$\text{where } L^{(t)}_{s^{(t)},v^{(t)}} = \sum_{r=1}^{p}\sum_{j=1}^{q}\left(s_j\tilde{\lambda}^{(t)}_{rv_j} - \lambda^{\star}_{rj}\right)^2$$

$$\text{so } min\Psi(\Lambda^{\star}, s, v) = \sum_{t=1}^{T}\sum_{r=1}^{p}\sum_{j=1}^{q}\left(s^{(t)}_j\tilde{\lambda}^{(t)}_{rv_j} - \lambda^{\star}_{rj}\right)^2$$

$$\text{subject to } s^{(t)}_j \in \{-1, 1\} \ , t = 1, \ldots, T \ ; j = 1, \ldots, q$$

$$\text{and } v^{(t)} \in T_q \ , t = 1, \ldots, T.$$

An important note about the minimization is that the reference loading matrix $\Lambda^{\star}$ is unknown and must be approximated via a recursive algorithm. The key idea for the solution of the sign and column switching of the RSP method was inspired by the solution of identifiability issues in Bayesian analysis of mixture models known as label switching (Stephens, 2000; Papastamoulis & Iliopoulos, 2010; Rodríguez & Walker, 2014; Papastamoulis, 2015).

The RSP methods main objective is the selection of the $(\Lambda^{\star}, s, v)$, so that they can minimize (2.1). The RSP methods achieve this in two parts. In the first part, which is called Reference-Loading Matrix Estimation RLME part, the RSP methods minimize $\Psi(\Lambda^{\star}, s, v)$ with respect to $\Lambda^{\star}$ for given values of (s, v). In the second part, which is called SP part, the RSP methods minimize $\Psi(\Lambda^{\star}, s, v)$ with respect to (s, v) given $\Lambda^{\star}$.

All of the RSP variants in the RLME part find the minimum of $\Psi(\Lambda^{\star}, s, v)$ with respect to $\Lambda^{\star}$ for given values of (s, v) to be equal to

$$\lambda^{\star}_{rj} = \frac{1}{T}\sum_{t=1}^{T} s^{(t)}_j \tilde{\lambda}^{(t)}_{rv^{(t)}_j}.$$

The differentiation between the RSP variants is in how each variant solves the SP part which is the minimization of $\Psi(\Lambda^{\star}, s, v)$ with respect to (s, v) given $\Lambda^{\star}$.

Rotation Sign permutation Exact

In the SP part of the RSP exact variant, the minimum of $\Psi(\Lambda^{\star}, s, v)$ with respect to (s, v) given $\Lambda^{\star}$ can be calculated by solving the assignment problem for all 2^q combinations of possible sign configurations of $s^{(t)} \in \{-1, 1\}^q$. Afterwards, the combination of (s, v) which is the overall minimum has to

be found. The calculation of $min_{v\in T_q}\{L_{s,v}^{(t)}\}$ individually for all of the 2^q sign configuration of s is a variety of the transport problem known as the assignment problem (Burkard & Dell'Amico, 2009).

The minimization problem of $min_{v\in T_q}\{L_{s,v}^{(t)}\}$ individually for all of the 2^q given sign matrices $S = diag(s_1, s_2, \ldots, s_q)$ is expressed as the following assignment problem

$$\min_{\delta_{ij}\in\{0,1\}\ ,\ i,j=1,2,\ldots q} \sum_{i=1}^{q}\sum_{j=1}^{q} \delta_{ij}c_{ij}$$

$$\text{subject to } \sum_{i=1}^{q} \delta_{ij} = 1\ \forall i = 1,\ldots,q$$

$$\text{subject to } \sum_{j=1}^{q} \delta_{ij} = 1\ \forall j = 1,\ldots,q$$

where C is the $q \times q$ cost matrix of the assignment problem and has the following form

$$c_{ij} = \sum_{r=1}^{p} \left(s_j\tilde{\lambda}_{rj} - \lambda_{ri}^{\star}\right)^2\ ,\ i,j = 1,\ldots,q$$

where δ_{ij} is the corresponding binary decision variables and has the following form

$$\delta_{ij} = \begin{cases} 1 & , \quad \text{if index } i \text{ is assigned to index } j \\ 0 & , \quad \text{otherwise} \end{cases}.$$

In order to solve the discrete combinatorial optimization problem, the RSP exact method uses the branch and bound technique (Little, Murty, Sweeney, & Karel, 1963). The algorithm for the RSP exact method is the following

Varimax-Rsp Exact Algorithm(correcting rotation,signed and permutation invariance to a MCMC sample)

Input : Simulated MCMC sample of factor loading$\left\{\Lambda^{(t)}, t = 1,\ldots,T\right\}$

Output : Reordered MCMC sample of factor loading$\left\{\Lambda^{\circ(t)}, t = 1,\ldots,T\right\}$

Step 1 : Varimax Rotation Step
for $t = 1$ **to** T
implement the Varimax rotation on $\Lambda^{(t)}$ and obtain the rotated Varimax loadings $\tilde{\Lambda}^{(\tau)}$
endfor

Step 2: Signed Permutation Step

Step 2.1 Initialization :
for all $t \in \{t = 1, \ldots, T\}$: Initialize $s^{(t)}$ and $v^{(t)}$
Proposed initialization
Set $s_j^{(t)} = 1$ and $v_j^{(t)} = j$ (identity permutation),$t = 1, \ldots, T$, $j = 1, \ldots q$
end

Step 2.2
Repeat

Step 2.2.1 RLME Step
set $\lambda^{\star}_{rj} = \frac{1}{T}\sum_{t=1}^{T} s_j^{(t)} \tilde{\lambda}^{(t)}_{rv_j^{(t)}}$, $r = 1, \ldots, p$,$j = 1, \ldots, q$
where $\tilde{\lambda}^{(t)}_{rv_j^{(t)}}$ are the Varimax rotated loadings

Step 2.2.2 SP step
for each iteration $t = 1, \ldots, T$

Step A.1
for each $s = (s_1, \ldots, s_q) \in \{-1, 1\}^q$
Find the permutation $v(s)$ that minimizes $\left\{L^{(t)}_{s,v} : v \in T_q\right\}$ by solving the assignment problem
endfor

Step A.2
set$(s^{(t)}, v^{(t)}) = argmin_{s,v}\left\{L^{(t)}_{s,v(s)} : s \in \{-1, 1\}^q\right\}$

endfor

Until no improvement in $\Psi(\Lambda^{\star}, s, v) = \sum_{t=1}^{T} L^{(t)}_{s^{(t)},v^{(t)}}$ is observed

End of algorithm

The RSP Exact variant of the RSP method deals with the SP part of the algorithm(The minimization of $\Psi(\Lambda^{\star}, s, v)$ with respect to (s, v) given $\Lambda^{\star}$)

by solving 2^q assignment problems and finding the overall minimum. This method is faster than solving all possible combinations of sign and column permutation($2^q q!$ different combinations). However, having a model with number of factors more than 10, the RSP exact variant becomes computationally non-efficient. The other two variants based on simulation annealing were created in order to solve this issue.

Full Simulation Annealing Rotation Sign permutation

The Full simulation annealing RSP is a RSP variant based on simulation annealing. In order to minimize $\Psi(\Lambda^\star, s, v)$ with respect to (s, v) given $\Lambda^\star$(SP part), the method operates as follows. It proposes a candidate state $(s^\star, v^\star)$ created by randomly changing one element of the current sign vector s and simultaneously permuting two randomly selected indexes in v. This procedure is repeated B times for each MCMC iteration controlled by a temperature T_b which is constantly cooling down and by doing this, the acceptance probability of each iteration $i = 1, \dots B$ is controlled. The candidate state is either approved as the new state or disapproved and the old state becomes the new state. The Full simulation annealing RSP SP step algorithm is the following.

Varimax-Rsp Full Simulated Annealing Algorithm(correcting rotation,signed and permutation invariance to a MCMC sample)

Step 2.2.2 SP step
for each iteration $t = 1, \dots, T$

Step A.1
Set initial values $(s^{(t,0)}, v^{(t,0)}) = (s^{(t)}, v^{(t)})$
and let $L^{(t,0)} = L^{(t)}_{s^{(t)}, v^{(t)}}$

Step A.2
for $b = 1$ to B

(a) Propose a candidate state $(s^\star, v^\star)$
Randomly switch the sign of one index in $s^{(t,b-1)}$, and
Permute the values of a randomly selected pair of indices in $v^{(t,b-1)}$

(b) Compute$L^\star = L^{(t)}_{s^\star, v^\star}$

(c) Set $(s^{(t,b)}, v^{(t,b)}) = (s^\star, v^\star)$ and $L^{(t,b)} = L^\star$ with probability

$$P\left((s^{(t,b-1)}, v^{(t,b-1)}) \to (s^\star, v^\star)\right) = \begin{cases} \exp\left(-\frac{L^\star - L^{(t,b-1)}}{T_b}\right) & \text{If } L^\star - L > 0 \\ 1 & \text{If } L^\star - L \leq 0 \end{cases}$$

Otherwise set $(s^{(t,b)}, v^{(t,b)}) = (s^{(t,b-1)}, v^{(t,b-1)})$ and $L^{(t,b)} = L^{(t,b-1)}$

endfor

Step A.3
Set $(s^{(t)}, v^{(t)}) = (s^{(t,B)}, v^{(t,B)})$

endfor

Until no improvement in $\Psi(\Lambda^\star, s, v) = \sum_{t=1}^{T} L^{(t)}_{s^{(t)}, v^{(t)}}$ is observed

End of algorithm

Partial Simulation Annealing Rotation Sign permutation

The partial simulation annealing RSP is a RSP variant which is a hybrid between the exact variant and the fully simulation annealing one. In order to minimize $\Psi(\Lambda^\star, s, v)$ with respect to (s, v) given $\Lambda^\star$(SP part), the method operates with the following way. It proposes a candidate state $(s^\star, v^\star)$ created by randomly generating a candidate sign configuration $s^\star$ from the current state, and afterwards, deterministically identifying the permutation $v^\star$ minimizing $\left\{L^{(t)}_{s^\star, v} : v \in T_q\right\}$ given $s^\star$. This procedure is repeated B times for each MCMC iteration controlled by a temperature T_b which is constantly cooling down and by doing this, the acceptance probability of each iteration $i = 1, \ldots B$ is controlled. The candidate state is either approved as the new state or disapproved and the old state becomes the new state. The partial simulation annealing RSP SP step algorithm is the following

Varimax-Rsp Partial Simulated Annealing Algorithm(correcting rotation,signed and permutation invariance to a MCMC sample)

Step 2.2.2 SP step
for each iteration $t = 1, \ldots, T$

Step A.1

Set initial values $(s^{(t,0)}, v^{(t,0)}) = (s^{(t)}, v^{(t)})$
and let $L^{(t,0)} = L^{(t)}_{s^{(t)},v^{(t)}}$

Step A.2

for $b = 1$ to B

(a) Propose a candidate state $(s^\star, v^\star)$
Obtain $s^\star$ by randomly switching the sign of one index in $s^{(t,b-1)}$
Obtain the permutation $v\star = v(s^\star)$ that minimizes $\left\{L^{(t)}_{s^\star,v} : v \in T_q\right\}$ by solving the assignment problem

(b) Compute$L^\star = L^{(t)}_{s^\star,v^\star}$

(c) Set $(s^{(t,b)}, v^{(t,b)}) = (s^\star, v^\star)$ and $L^{(t,b)} = L^\star$ with probability

$$P\left((s^{(t,b-1)}, v^{(t,b-1)}) \to (s^\star, v^\star)\right) = \begin{cases} \exp\left(-\frac{L^\star - L^{(t,b-1)}}{T_b}\right) & \text{If } L^\star - L > 0 \\ 1 & \text{If } L^\star - L \leq 0 \end{cases}$$

Otherwise set $(s^{(t,b)}, v^{(t,b)}) = (s^{(t,b-1)}, v^{(t,b-1)})$ and $L^{(t,b)} = L^{(t,b-1)}$

endfor

Step A.3

Set $(s^{(t)}, v^{(t)}) = (s^{(t,B)}, v^{(t,B)})$

endfor

Until no improvement in $\Psi(\Lambda^\star, s, v) = \sum_{t=1}^{T} L^{(t)}_{s^{(t)},v^{(t)}}$ is observed

End of algorithm

An important note about the last two RSP variants is that one iteration of the partial simulation annealing is computationally more demanding than one iteration of the full simulation annealing. However, the partial simulation annealing requires fewer iterations in order to converge to the target distribution in contrast to the full simulation annealing.

2.5 MatchAlign

MatchAlign is a post-hoc method created by Poworoznek et al. (2021) in order to solve the identification issue regarding the factor loadings matrix. The

MatchAlign method can be employed with any model. For the generation of the MCMC samples of the parameters of the factor model regarding the MatchAlign method two different models will be used. The first one is by utilizing the Dirichlet-Laplace shrinkage prior, proposed by Bhattacharya, Pati, Pillai, and Dunson (2015). The second one is by employing the Multiplicative Gamma Shrinkage prior, proposed by Bhattacharya and Dunson (2011). Before proceeding with the description of the MatchAlign method, the two models will be presented.

Dirichlet-Laplace Shrinkage Prior

Dirichlet-Laplace shrinkage prior is a Bayesian shrinkage method based on the idea of Global Local (GL) mixtures of Gaussians with the attribute that the entire posterior distribution concentrates at the optimal rate (Bhattacharya et al., 2015), which leads to depiction of uncertainty in estimates of parameters and functions of parameters. Before proceeding with the definition of the Dirichlet-Laplace shrinkage prior some concepts have to be defined, which are prerequisite in order to understand the prior.

The Global Local mixtures of Gaussians shrinkage priors family are the priors which have the following form

$$\theta_j \sim N(0, \psi_j t) \ , \ \psi_j \sim f \ , \ t \sim g$$

where t controls the global shrinkage towards the origin and ψ_j is the local scales which enable variation in the magnitude of shrinkage.

By integrating out the local scales ψ_j, the Global local mixtures of Gaussians can be rewritten as a global scale mixture of kernel $K(\centerdot)$ in the following form

$$\theta_j \sim K(\centerdot, t) \ , \ t \sim g$$
$$\text{where } K(x) = \int \psi^{-\frac{1}{2}} \Phi_0(\frac{x}{\sqrt{\psi}}) g(\psi) d\psi$$
$$\text{and } K(x,t) = t^{-\frac{1}{2}} K(\frac{x}{\sqrt{t}})$$

where $K(x)$ is a symmetrical density in $\mathbb{R}$ and Φ_0 denotes the standard normal density in $\mathbb{R}$.

The idea in Dirichlet-kernel priors is to substitute the single global scale t in the global scale mixture of kernel with a vector of scales $(\varphi_1 t, \varphi_2 t, \ldots, \varphi_p t)$ with the restriction that the $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_p)$ belongs in the $p-1$ dimensional simplex $S^{p-1} = \left\{ x = (x_1, x_2, \ldots, x_p)^T : x_j \geq 0 \ , \ \sum_{j=1}^{p} x_j = 1 \right\}$ and is allocated a Dirichlet prior $\varphi \sim Dir(a, a, \ldots, a)$. The Dirichlet-kernel priors have the following form

$$\theta_j | \varphi_j, t \sim K(\centerdot, \varphi_j, t) \ , \ \varphi \sim Dir(a, a, \ldots, a)$$

K can be any unimodal density which is symmetric around zero having exponential or heavy tails. An important note is that by choosing a kernel which can be rewritten as a scale mixture of normals the computations become more efficient.

By choosing the Laplace distribution as the kernel, we have the Dirichlet-Laplace prior. Any parameter following this prior will be denoted as $\theta | t \sim DLa(t)$. The Dirichlet-Laplace prior objective is to appear like the joint distribution of θ under a two-component mixture prior. The Dirichlet-Laplace prior has the following form

$$\theta_j | \varphi_j, t \sim DE(\varphi_j, t) \ , \ \varphi \sim Dir(a, a, \ldots, a) \ , \ t \sim Gamma(pa, 1/2)$$

where $DE(t)$ denotes a Laplace or double Exponential distribution with density $f(y) = (2t)^{-1} e^{-\frac{|y|}{t}}$ for $y \in \mathbb{R}$.

The parameter t is of great importance due to its attribute to regulate the tails of the marginal distribution of θ_j. The form of the marginal distribution of $\theta_j | t$ for any $a < 1$ is unbounded with a singularity at 0.

The data augmented Gibbs sampler should be used in order to generate Dirichlet-Laplace posterior samples but, before proceeding with the explanation of the algorithm some concepts have to be define.

The Dirichlet-Laplace prior can be expressed in the following form

$\theta_j \sim N(0, \psi_j \varphi_j^2 t^2)$, $\psi_j \sim Exp(1/2)$, $\varphi \sim Dir(a, \ldots, a)$, $t \sim Gamma(pa, 1/2)$.

Bhattacharya and Dunson (2011) proved that the joint posterior of $\varphi|\theta$ and $(\frac{T_1}{T}, \frac{T_2}{T}, \ldots, \frac{T_p}{T})$ have the same distribution when T_j are independently distributed and they follow a three parameter generalized inverse-Gaussian(gIGaussian) distribution and $T = \sum_{j=1}^{p} T_j$.

A variable will follow the three parameter generalized inverse-Gaussian (gIGaussian) distribution $y \sim gIGaussian(\lambda, p, x)$ if the density of y is $f(y) \propto y^{(\lambda-1)} e^{(-\frac{1}{2})(px+\frac{x}{y})}$ for $y > 0$.

In order to generate samples from a Dirichlet-Laplace posterior using an augmented Gibbs sampler the following steps should be followed

Step 1 Sample $\theta|\psi, \varphi, \tau, y$ by generating θ_j independently from a normal distribution $\theta_j \sim N(\mu_j, \sigma_j^2)$ with mean parameter equal to $\mu_j = (1 + \frac{1}{(\psi_j \varphi_j^2 t^2)})^{-1} y$ and with variance parameter equal to $\sigma_j^2 = (1 + \frac{1}{(\psi_j \varphi_j^2 t^2)})^{-1}$.

Step 2 Sample the conditional posterior of $\psi|\varphi, t, \theta$ by generating samples in a block by independently sampling $\tilde{\psi}_j|\varphi, \theta$ from an inverse Gaussian $IGaussian(\mu_j, \lambda)$ with $\mu_j = \varphi_j \frac{t}{|\theta_j|}$, $\lambda = 1$ and setting $\psi_j = \frac{1}{\tilde{\psi}_j}$.

Step 3 Generate the conditional posterior of $t|\varphi, \theta$ from a three parameter generalized inverse-Gaussian(gIGaussian) distribution $gIGaussian(\lambda - p, 1, 2\sum_{j=1}^{p} \frac{|\theta_j|}{\varphi_j})$.

Step 4 In order to sample from $\varphi|\theta$, generate $(T_1, T_2, \ldots, T_p)$ independently from a three parameter generalized inverse-Gaussian(gIGaussian) distribution $T_j \sim gIGaussian(a - 1, 1, 2|\theta_j|)$ and set $\varphi_j = \frac{T_j}{T}$ where $T = \sum_{j=1}^{p} T_j$.

In this thesis the Dirichlet-Laplace prior is used in the row of the factor loadings matrix Λ and an inverse-Gamma prior used on the diagonal elements of the variance covariance matrix of the idiosyncratic errors Σ with parameters $(\frac{1}{2}, \frac{1}{2})$. In mathematical notation the previous sentence can be expressed as

$$\lambda_{ik}|\varphi_{ik}, t_i \sim DE(\varphi_{ik} t_i) \; k = 1, \ldots, K$$

$$\varphi_i \sim Dir(a, \ldots, a)\ t_i \sim G(Ka, \frac{1}{2})$$

$$\sigma_{ii} \sim IG(\frac{1}{2}, \frac{1}{2})$$

where $i = 1, \ldots, p$,$\varphi_i = (\varphi_{i1}, \ldots, \varphi_{iK})$, K is the maximum number of factors and a a hyper prior.

Multiplicative Gamma Process Shrinkage Prior

Multiplicative Gamma Process shrinkage prior permits the employment of infinitely number of factors on a parameter expanded factor loadings matrix. Having a parameter expanded factor loadings matrix makes the constrains in the ordering of the factors pointless. Consequently, Multiplicative Gamma process shrinkage prior is free from order dependences. Furthermore, the factor loadings in the factor loadings matrix are progressively shrunk to zero as the column index grows. The central principle of the method derives from the fact that for inference or prediction on the covariance matrix, the identifiability of the factor loadings is redundant. Consequently, the requirement for a unique decomposition is unnecessary.

Multiplicative Gamma Process Shrinkage Prior has the following prior specification

$$\lambda_{jh}|\Phi_{jh}, t_h \sim N(0, \Phi_{jh}^{-1} t_h^{-1})\ ,\ \Phi_{jh} \sim Gamma(\frac{3}{2}, \frac{3}{2})\ ,\ t_h = \prod_{l=1}^{h} \delta_l$$

$$\delta_1 \sim Gamma(a_1, 1)\ ,\ \delta_l \sim Gamma(a_2, 1)\ l \geq 2\ ,\ \sigma_j^{-2} \sim Gamma(a_\sigma, b_\sigma)\ ,\ j = 1, \ldots, p$$

where δ_l , $l = 1, 2, \ldots, \infty$ are independent. t_h denotes the global shrinkage parameter for the h-th column and Φ_{jh} denotes the local shrinkage parameters for the objects in the h-th column. The prior for the diagonal objects of Σ is the conventional Inverse Gamma prior. The prior of the factor loadings matrix makes the objects in each column reduce in proportion flexibly as the column index grows. Furthermore, the prior uses a parameter expanded factor loading matrix and does not force any regulations on the factor loading objects. An important note about the prior is that under the regulation that $a_2 > 1$, t_h stochastically increase, leading to greater shrinkage as the column index grows. Additionally by employing only the global shrinkage parameter, the prior will over-shrink the non-zero factor loading. As a result, in order for the prior to operate correctly, we need to employ both the global

shrinkage parameter and the local shrinkage parameters.

In order to generate a sample from the posterior distribution of the Multiplicative Gamma Process shrinkage prior a Gibbs sampler should be used with the following steps as described in Bhattacharya and Dunson (2011)

Step 0 Truncate the loading matrix in order to have $q^\star << p$ columns and iterate the following steps(1-6)

Step 1 Generate λ_js from

$$\pi(\lambda_j|-) \sim N_{q^\star}\left((D_j^{-1} + \sigma_j^{-2}F^TF)^{-1}F^T\sigma_j^{-2}Y^{(j)}, (D_j^{-1} + \sigma_j^{-2}F^TF)^{-1}\right)$$

where λ_j^T denotes the j-th row of $\Lambda_{q^\star}$, $F^T = (F_1, F_2, \ldots, F_n)$,$D_j^{-1} = diag(\Phi_{j1}t_1, \ldots, \Phi_{jq^\star}t_{q^\star})$ and $Y^{(j)} = (Y_{1j}, Y_{2j}, \ldots, Y_{nj})^T$ for $j = 1, 2, \ldots, p$.

Step 2 Generate σ_j^{-2} from

$$\pi(\sigma_j^{-2}|-) \sim G\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(Y_{ij} - \lambda_j^T F_i)^2\right),\ j = 1, \ldots, p.$$

Step 3 Generate F_i from

$$\pi(F_i|-) \sim N_{q^\star}\left((I_{q^\star} + \Lambda_{q^\star}^T\Sigma^{-1}\Lambda_{q^\star})^{-1}\Lambda_{q^\star}^T\Sigma^{-1}Y_i\ ,\ (I_{q^\star} + \Lambda_{q^\star}^T\Sigma^{-1}\Lambda_{q^\star})^{-1}\right),\ i = 1, \ldots, n.$$

Step 4 Generate Φ_{jh} from

$$\pi(\Phi_{jh}|-) \sim G\left(\frac{v+1}{2}\ ,\ \frac{v + t_h\lambda_{jh}^2}{2}\right)$$

where $v = 3$.

Step 5 If $h = 1$ generate δ_1 from

$$\pi(\delta_1|-) \sim G\left(a_1 + \frac{pq^\star}{2}\ ,\ 1 + \frac{1}{2}\sum_{l=1}^{q^\star}t_l^{(1)}\sum_{j=1}^{p}\Phi_{jl}\lambda_{jl}^2\right)$$

else if $h \geq 2$ generate δ_h from

$$\pi(\delta_h|-) \sim G\left(a_2 + \frac{p}{2}(q^\star - h + 1)\ ,\ 1 + \frac{1}{2}\sum_{l=h}^{q^\star} t_l^{(h)} \sum_{j=1}^{p} \Phi_{jl}\lambda_{jl}^2\right)$$

where $t_l^{(h)} = \prod_{t=1,t\neq h}^{l} \delta_t$ for $h = 1, 2, \ldots, p$.

Step 6 Generate a_1 and a_2 employing a Metropolis-Hastings step within the Gibbs sampler.

Selecting the number of factors using an adaptive algorithm

Bhattacharya and Dunson (2011) suggested an adaptive Gibbs sampler (Roberts & Rosenthal, 2007) which simultaneously generates a sample from the posterior distribution of the parameter and also selects the number of factors. The proposed adaptive Gibbs sampler can be summarized in the following steps:

Step 0 Initiate with a conservative speculation $\tilde{q}$ of $q^\star$ where $q^\star$ denotes the effective number of factors. The effective number of factors is the number of factors about which the introduction of additional factors will not contribute more information to the model. Bhattacharya and Dunson (2011) used the value $5\log(p)$ as the starting one for $\tilde{q}$ in their simulated experiments. p denotes the number of observed variables.

Step 1 Initiate the above Gibbs sampler with number of columns in the factor loadings matrix equal to $\tilde{q}$ and also select (a_0, a_1). An important note about the selection of (a_0, a_1) is that they must be selected in such a way that at the beginning of the chain the adaptation happens approximately every 10 iterations but after some iterations, it will happen exponentially more frequently. Bhattacharya and Dunson (2011) in their simulated experiments set $a_0 = -1$ and $a_1 = -5 * 10^{-4}$.

Step 1 A Generate a sequence of random variable u_t from a uniform distribution $U(0, 1)$.

Step 1 B If the random variable u_t is less than the adapt probability $u_t < \exp(a_0 + a_1 t)$ at some iteration t, check the generated factor loadings matrix generated at that iteration t for columns having all their objects inside a pre-specified neighbourhood near 0. Bhattacharya and Dunson (2011) in their simulated experiments

checked all columns for factor loadings having absolute value less than 10^{-4}. Then, discard all the columns with all their objects inside a pre-specified neighbourhood near 0. If such column does not exist, add an extra column and generate its factor loadings using the prior specified in the previous Gibbs sample. Additionally, if the random variable u_t is less than the adapt probability, set the effective number of factors at iteration t to be equal to $q^{\star(t)} = \tilde{q}^{(t)} - m^{(t)}$. Where $m^{(t)}$ denotes the number of columns having all their objects inside a pre-specified neighbourhood near of 0 in the t-th iteration and $\tilde{q}^{(t)}$ is the truncation level at the t-th iteration.

Step 2 After the termination of the Gibbs sampler and after the burn in, choose as the appropriate number of factor the median of the effective number of factors $median(q^{\star(t)})$ and use the generated samples $\Omega^{(T)} = \Lambda^{(t)}_{\tilde{q}^{(t)}} \Lambda^{(t)T}_{\tilde{q}^{(t)}} + \Sigma^{(t)}$ of the approximated marginal posterior distribution of Ω in order to inference on Ω.

An important note about the correct implementation of the adaptive Gibbs sampler is that in order for the algorithm to operate correctly, the proposed $\tilde{q}$ must be greater than the true number of factors. Additionally, this method, increasingly shrinking the factor loadings matrix while the column index growing, deals with the problem of Factor Splitting appearing in some other shrinkage priors. The problem of Factor Splitting is that the columns of factor loading matrix do not have columns having all their objects in some neighbour near 0, even when the number of columns is greater than the true number of factors. Finally, the convergence of the chain to the posterior distribution of the parameters is ensured as indicated by Roberts and Rosenthal (2007).

MatchAlign

MatchAlign is a post-hoc method created to solve the identification issue regarding the factor loadings matrix inspired by the method proposed by Papastamoulis and Ntzoufras (2022a) and Marin, Mengersen, Robert, Dey, and Rao (2005). The MatchAlign method can be summarized into three steps. The first step is the orthogonalization step. In this step, the method uses an orthogonalization procedure, specifically the Varimax rotation (Kaiser, 1958). Nevertheless, any other orthogonalization procedure could be implemented (For example Oblique rotation (Thurstone, 1931)). The second step

is the selection of the pivot matrix. The third step is a greedy maximization procedure in which the method pairs the columns of each posterior sample with the pivot matrix and uses the pairs to align the samples afterwards. After a brief overview of the method each step is to be explained in detail.

Denoting as $\left\{\Lambda^{(t)}\ , t = 1, \ldots, T\right\}$ the posterior samples of the factor loadings matrix Λ. In the orthogonalization step, the method deal with generic invariance between the posterior samples $\Lambda^{(t)}$ and enables the posterior samples to be depicted in an interpretable form. For the sake of achieving that, the method employs an orthogonalization procedure to each sample $\Lambda^{(t)}$. The method uses the Varimax rotation although Oblique rotation can also be used and produce valid representations when correlated factors exist. The main issue with the Oblique rotation is that the oblique rotated matrices may not be comparable across matrices (Poworoznek et al., 2021). Finally, $\tilde{\Lambda}$ will denote the factor loadings matrix after applying the Varimax rotation.

In the second step, the method must select a pivot matrix and employ it as a proxy of the true factor loadings matrix Λ. Poworoznek et al. (2021) used the Condition Number in order to select a pivot matrix having the attribute of being centralized in the distribution of a column statistic. The Condition Number, according to Poworoznek et al. (2021), is a proxy of distinctive knowledge contained in each column having the following form

$$k^{(t)} = k(\Lambda^{(t)}) = \frac{\sigma_{max}(\Lambda^{(t)})}{\sigma_{min}(\Lambda^{(t)})}$$

where $\sigma_{max}(\Lambda^{(t)})$ denotes the largest singular value of the matrix $\Lambda^{(t)}$ and $\sigma_{min}(\Lambda^{(t)})$ denotes the smallest singular value of the matrix $\Lambda^{(t)}$. The pivot matrix is selected as the matrix with the median condition number on condition that the condition number does not approach infinity. In the occasion that the condition number approaches infinity, the pivot matrix must be selected as the matrix with the median largest singular value.

In the third step of the method, the identification issue in the factor loading matrix caused by label and sign switching is solved. In order to achieve that, the matrices aligning the samples and consequently minimizing the following loss function must be identified.

$$minimize_{Q^{(t)},S^{(T)},t=1,\dots T} = \sum_{t=1}^{T} ||\tilde{\Lambda}^{(t)} Q^{(t)} S^{(t)} - \Lambda^P||_F$$

where Q denotes a $q \times q$ permutation matrix, $S = diag(s_1, s_2, \dots, s_q)$ with $s_j \in \{-1, 1\}$ a signs vector and $P = SQ$ a signed permutation matrix as defined in Papastamoulis and Ntzoufras (2022a) and as explained in (2.4). Additionally, $||\ ||_F$ denotes the Frobenious norm, $\Lambda^P = \left[c_1^P, c_2^P, \dots, c_q^P\right]$ the pivot matrix and $\tilde{c}_j^{(t)}$ the j-th column of $\tilde{\Lambda}^{(t)}$.

Taking into account that the pivot matrix and the generated samples are penalized differently by the loss function, the sign changes and columns reordering of $(\tilde{c}_1^{(t)}, \tilde{c}_2^{(t)}, \dots, \tilde{c}_q^{(t)})$ matching the pivot matrix must be identified so that the loss function can be minimised. In order to achieve that the method repetitively calculates the L_2 normed difference of the columns of the matrices $\tilde{\Lambda}^{(t)}$ and the columns of the matrices Λ^P and its negative $-\Lambda^P$. This iterative procedure was inspired by the method of Marin et al. (2005). This minimization procedure can be summarized in the following steps : 1) Calculate the L_2 normed difference between the column of $\tilde{\Lambda}^{(t)}$ having the biggest norm and the columns $\left[c_1^P, -c_1^P, c_2^P, -c_2^P, \dots, c_q^P, -c_q^P\right]$. 2) Identify the column c_j^P or its negative $-c_j^P$ for $j = 1, 2, \dots, q$ minimizing the L_2 norm difference with the column of the matrix $\tilde{\Lambda}^{(t)}$ with the biggest norm (match them). 3) Remove the j-th column and its negative $(-c_j^P, c_j^P)$. 4) Proceed with calculating the L_2 normed difference between the column of $\tilde{\Lambda}^{(t)}$ having next biggest norm after the previous one and the columns $\left[c_1^P, -c_1^P, c_2^P, -c_2^P, \dots, c_{j-1}^P, -c_{j-1}^P, c_{j+1}^P, -c_{j+1}^P, \dots, c_q^P, -c_q^P\right]$. 5) Continue this iterative procedure until there are not any columns left. On the occasion that the prior distribution of Λ is a shrinkage prior with the attribute that the shrinkage increases as the column index grows as in the prior distribution proposed by Bhattacharya and Dunson (2011). Instead of using the column with the biggest norm, the first column will be used. Afterwards, advance to the next column according to column order.

An issue regarding the MatchAlign minimization of the loss function procedure is the duplication. According to Poworoznek et al. (2021) duplication occurs when one $\tilde{c}_j^{(t)}$ is minimally distant from many c_h^P. The main reason for duplication is sampling noise. Finally, the consequence of duplication is biased ergodic summaries of the parameters.

An important note about the MatchAlign method is that it minimizes the difference of the pivot matrix and the generated samples iterative column

by column. By doing so it does not minimize the difference between them globally. This results in some columns centred near 0_p to be mislabelled. However, this does not prejudice the ergodic summaries of the parameters as long as the the mislabelled columns have small norm difference with the pivot matrix columns.

Chapter 3

Comparison of the different models in synthetic data sets

The main objective of the third chapter of this thesis is to compare the efficiency of the different models by using synthetic data. The comparison will be made in the following fields : 1) How much time each method requires in order to solve rotational ambiguities and label switching. 2) The number of factors selected by the model in comparison with the true number of factors (The true number of factors is known because the datasets are synthetic). 3) The number of variables which are allocated correctly (we know which variables must be allocated together to the same factors because the datasets are synthetic). 4) A metric evaluating the performance of solving rotational ambiguities and label switching.

For the generation of the synthetic data sets the method proposed by Papastamoulis (2018, 2020) will be used. A brief description of this method is the following. The method was originally conceived in order to create synthetic data sets for finite mixture of factor analyzers models. However, by setting the number of clusters into 1, is able to simulate data sets for factor analysis models. The simulated observed variables are generated from the following multivariate normal distribution

$$X_j \sim N_p(\mu, \Lambda\Lambda^T + \Sigma),\ j = 1, \ldots, n.$$

For a pre-specified sample size n, number of factors q, number of observed variables p, grid of values for the means of the factor loadings φ, a vector with the values of the diagonal elements of the inverse covariance matrix and the standard deviation of the factor loadings ρ, the true values of the

parameters of interest are created as follows

$$\mu = \begin{cases} 20\sin(\frac{r-1}{p-1}\pi) \text{ with probability } \frac{1}{3} \\ 20\cos(\frac{r-1}{p-1}\pi) \text{ with probability } \frac{1}{3} \\ -40\cos(\frac{r-1}{p-1}2\pi) \text{ with probability } \frac{1}{3} \end{cases}$$

$$\Lambda^T = \left|\begin{array}{cccc||cccc|c|cccc||ccc|} \star & \star & \dots & \star & \square & \square & \dots & \square & \dots & \square & \square & \dots & \square & \square & \dots & \square \\ \square & \square & \dots & \square & \star & \star & \dots & \star & \dots & \square & \square & \dots & \square & \square & \dots & \square \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \square & \square & \dots & \square & \square & \square & \dots & \square & \dots & \star & \star & \dots & \star & \square & \dots & \square \end{array}\right|$$

A star $\star$ denotes independent depictions from a random variable following a normal distribution $N(\varphi, \rho^2)$ and a square $\square$ independent depictions from a random variable following the standard normal distribution $N(0, 1)$. Every block matrix with a $\star$ element has q rows and $\lfloor \frac{p}{q} \rfloor$ columns($\lfloor \rfloor$ denotes the floor function). On the occasion that $\frac{p}{q}$ is a rational number, the rest of the columns $p - \lfloor \frac{p}{q} \rfloor$ will be set to square $\square$ (analogous to the last block of Λ^T).

The metric which will be used in order to evaluate the performance of solving rotational ambiguities and label switching by each method was proposed by Poworoznek et al. (2021). The metric is the normed dissimilarity of the posterior mean of the covariance matrix from the covariance matrix which is estimated by the posterior mean of the factor loadings matrix Λ after rotational ambiguities and label switching problems being solved. The metric has the following form

$$||\overline{\Lambda\Lambda^T} - \bar{\Lambda}_\star \bar{\Lambda}_\star^T||_F$$

where $\overline{\Lambda\Lambda^T} = \frac{1}{T}\sum_{t=1}^{T} \Lambda^{(t)}(\Lambda^{(t)})^T$ and $\bar{\Lambda}_\star = \frac{1}{T}\sum_{t=1}^{T} \Lambda_\star^{(t)}$ denotes the posterior mean of the factor loadings matrices Λ after the method has solved rotational ambiguities and label switching problems. The metric has the following interpretation: when the normed difference is small, there is a indication that the method has solved the rotational ambiguities and label switching problems sufficiently, otherwise, the solution is not sufficient.

In order to examine if the chain has reached converges, the method proposed by Heidelberger and Welch (1981) will be used. For more information about the convergence diagnostic the reader is referred to the appendix of this

thesis. Furthermore, because the Heidelberger and Welch convergence diagnostic execute multiple comparisons, the Bonferroni correction (Bonferroni, 1936) will be applied in order to ensure that the possibility of incorrectly rejecting the null hypothesis is very low.

Whether, an observed variable change scale will have an after effect on the covariance of the observed variables and conclusively on the estimation of the parameters of factor model and their interpretation (Bartholomew et al., 2011). In order to prevent that, we standardize our observed variables before initiating our analysis. The observed variables are standardized in the following way

$$Z_i = \frac{X_i - \bar{X}_i}{\sigma_i}$$

where Z_i is the standardized observed variable, X_i is the observed variable, $\bar{X}_i$ is the mean of the observed variable and σ_i is the standard deviation of the observed variable.

The simulation of the synthetic data sets will follow the following three scenarios in order to evaluate the performance of the methods : 1) The first scenario will have sample size n equal to 1000, number of observed variables p equal to 9. The variance will be the same for all the idiosyncratic errors and equal to 100 and the number of true factors q equal to 3. 2) The second scenario will have sample size n equal to 1000, number of observed variables p equal to 27, variance of the idiosyncratic errors following a Uniform distribution $U(100, 1000)$ and number of true factors q equal to 9. 3) The third scenario will have sample size n equal to 500, number of observed variables p equal to 120, variance of the idiosyncratic errors following a Uniform distribution $U(100, 1000)$ and the number of true factors q equal to 20. For all three scenarios the grid of values for the mean of the factor loadings will be $(-30, -25, -20, 20, 25, 30)$ and the standard deviation of the factor loadings will be equal to 0.1. The first two scenarios will be implemented 10 times. The summary results will consist of the average of each outcome regarding the fields which have been mentioned at the beginning of this chapter. (For example, the average time required by a method to solve the rotational ambiguities and label switching problems). The third scenario will be implemented only one time because it is computationally demanding.

The association between the simulated observed variables and the factors in each of the three scenarios will have the following form: Upon the im-

plementation of the first scenario, the observed variables are produced in such a way that they can be divided in triads. In particular, the first three variables (123) are highly correlated with each other and low correlated with the other observed variables. Moreover, the rest of the triads $(456, 789)$ are subject to the same rationale. Consequently, as each observed variable in each triad is highly correlated to each other in the same triad, they must be allocated together to the same factor. As for the second scenario, the same logic is applied. However, in this case there are 9 triads in total. Finally, regarding the third scenario, triads are replaced with sextets $(123456, 789101112, \ldots, 115116117118119120)$ being subject to the same rationale as the previous scenarios.

All coding was conducted on the R programming language (Team, 2021). As already mentioned for the generation of the MCMC samples regarding the following methods : (RSP Exact,RSP Full Simulation Annealing, RSP Partial Simulation Annealing,WOP,OP) the MCMCpack ("MCMCpack", n.d.) package of R will be used with number of iterations equal to 1000000, burn in equal to 400000, thinning equal to 300. For the generation of the MCMC samples regarding the PLT method the MCMCpack ("MCMCpack", n.d.) package of R will be employed with number of iterations equal to 1000000, burn in equal to 400000, thinning equal to 300. For the generation of the MCMC samples for the MatchAlign method with the priors (Dirichlet-Laplace Shrinkage Prior, Multiplicative Gamma Shrinkage Prior) specified in the second chapter of this thesis the infinitefactor (Poworoznek, 2020) package of R will be used with number of iterations equal to 1000000, burn in equal to 400000, thinning equal to 300 and initial value for the number of factor for the Multiplicative Gamma Shrinkage Prior equal to $5 \log_{10} p$ as suggested by the authors in Bhattacharya and Dunson (2011). For the generation of the MCMC samples for the BEFA method the BayesFM (Piatek, 2021) package of R will be used with number of iterations equal to 1000000 and burn in equal to 800000. Furthermore, for the methods which do not choose the number of factors, only three chains will be presented having the number of factors equal to $(q-1, q, q+1)$ where q is the true number of factors. Additionally, for the method RSP Full Simulation Annealing and RSP Partial Simulation Annealing 1000 and 250 Simulation Annealing iterations(sa) will be executed respectively. The factor.switching (Papastamoulis & Ntzoufras, 2022b) packages of R which will be used for the implementation of the solution of the rotational ambiguities and label switching by these methods: RSP Exact, RSP Full Simulation Annealing, RSP Partial Simulation Annealing, WOP and OP. Similarly, the infinitefactor (Poworoznek, 2020) package of R will be used for MatchAlign method. Moreover, the BayesFM (Piatek, 2021)

package of R will be used for BEFA method. The PLT method does not require a method to solve rotational ambiguities and label switching because of the restrictions implemented in the generation of the MCMC samples. For the generation of the synthetic data sets the FabMix (Papastamoulis, 2018, 2020) package of R will be used.

The prior specifications for the models producing the posterior samples is the following

The prior distributions for the generation of the posterior samples by the Gibbs sampler of the MCMCpack are

For the rows of the factor loadings matrix Λ

$$\Lambda_{ij} \sim N(l_{0_{ij}}, {L_{0_{i,j}}}^{-}1) \text{ for } i = 1, \ldots, p,\ j = 1, \ldots, q$$

with value for $L^{-}_{0_{i,j}}1$ equal to 0 and for $l_{0_{ij}}$ also equal to 0. By setting $L^{-}_{0_{i,j}}1 = 0$ we are employing an improper prior.

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_i \sim IG(\frac{a_0}{2}, \frac{b_0}{2}),\ i = 1, \ldots, p$$

with value for a_0 and b_0 equal to 0.001.

The prior distributions for the generation of the posterior samples by the Positive Lower Triangular method are the following

For the non-diagonal elements of the factor loadings matrix

$$\Lambda_{ij} \sim N(0, \infty)\ \forall i \neq j\ i = 1, \ldots, p,\ j = 1, \ldots, q.$$

By setting the variance of the non-diagonal elements of the factor loadings matrix to ∞ we are employing an improper prior.

For the diagonal elements of the factor loadings matrix

$$\Lambda_{jj} \sim N(0, \infty)1(\Lambda_{ii} > 0)\ j = 1, \ldots, q.$$

This prior specification is a truncation of the normal distribution restricting the distribution to only positive values. By setting the variance of the diagonal elements of the factor loadings matrix to ∞ we are employing an improper prior

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_i^2 \sim IG(\frac{0.001}{2}, \frac{0.001}{2}),\ i = 1, \ldots, p.$$

The prior distributions for the generation of the posterior samples by the MatchAlign method with Dirichlet-Laplace prior are the following

For the rows of the factor loadings matrix Λ

$$\lambda_{ik}|\varphi_{ik}, t_i \sim DE(\varphi_{ik} t_i)\ k = 1, \ldots, K$$

$$\varphi_i \sim Dir(1/2, \ldots, 1/2)\ t_i \sim G(K\frac{1}{2}, \frac{1}{2})$$

where $i = 1, \ldots, p$,$\varphi_i = (\varphi_{i1}, \ldots, \varphi_{iK})$ and K is the maximum number of factors.

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_i \sim IG(\frac{1}{2}, \frac{1}{2}),\ i = 1, \ldots, p.$$

The prior distributions for the generation of the posterior samples by the MatchAlign method with Multiplicative Gamma Process Shrinkage Prior are the following

For the factor loadings matrix

$$\lambda_{jh}|\Phi_{jh}, t_h \sim N(0, \Phi_{jh}^{-1} t_h^{-1})\ ,\ \Phi_{jh} \sim Gamma(\frac{3}{2}, \frac{3}{2})\ ,\ t_h = \prod_{l=1}^{h} \delta_l$$

$$\delta_1 \sim Gamma(1, 1)\ ,\ \delta_l \sim Gamma(2, 1)\ l \geq 2\ ,\ j = 1, \ldots, p, l = 1, \ldots, h.$$

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_j^{-2} \sim Gamma(1, 0.3)\ ,\ j = 1, \ldots, p.$$

The prior distributions for the generation of the posterior samples by BEFA method are the following

The prior of the binary indicator matrix

$$Pr(\Delta_m = e_l|t_l) = t_l \ , \ \sum_{l=0}^{q} t_l = 1 \ , \ l = 1, 2, \ldots, q$$

where Δ_m , for $m = 1, 2, \ldots p$ is the m-th row of the binary indicator matrix Δ and e_k is the indicator vector

$$t^\star \sim Dir(\frac{1}{K}, \frac{1}{K} \ldots, \frac{1}{K}) \ , \ t_m = (t_{0m}, (1 - t_{0m})t_1^\star, (1 - t_{0m})t_2^\star, \ldots, (1 - t_{0m})t_q^\star)$$

$$\text{for } m = 1, 2, \ldots, p \text{ and } t_{0m} \sim Beta(2, 1)$$

where K is the maximum number of factors.

The prior distribution of the diagonal elements of the variance matrix Σ of the idiosyncratic errors is

$$\sigma_m^2 \sim IG\,(2, 1) \ \ m = 1, 2, \ldots, p.$$

The prior distribution of the only non-zero factor loadings in the m-th row is

$$\Lambda_{ml}^{\Delta}|\sigma_m^2 \sim N(0, 10\sigma_m^2), m = 1, \ldots, p.$$

The prior distribution of the covariance matrix Z with pre-specified S for the Inverse-Wishart prior is

$$Z \sim IW(K + 1, diag(1))$$

where K is the maximum number of factors and $diag(1)$ is a $q \times q$ diagonal matrix with all objects equal to 1.

The prior distribution of the covariance matrix Z with the Huang and Wand prior in S for the Inverse-Wishart prior is

$$Z \sim IW(K + 1, W) \ W = diag(w_1, w_2, \ldots, w_q)$$

$$\text{where } w_k \sim G(\frac{1}{2}, \frac{1}{4})$$

where K is the maximum number of factors.

Finally, the intermediate steps S are equal to

$$S = 1 + \Phi$$

$$\text{where } \Phi \sim Poisson(4).$$

3.1 Synthetic data scenario 1

Table 3.1: Summarized results for synthetic data scenario 1

Model and identification method	Time to solve identification issues	Factors by model	Variables loaded correctly	Metric	Time to simulate MCMC samples
PLT 2 q	-	2	6	-	15.108 mins
PLT 3 q	-	3	9	-	17.201 mins
PLT 4 q	-	3	9	-	31.992 mins
RSP exact 2 q	8.821 secs	2	6	0.00535	24.324 mins
RSP exact 3 q	18.491 secs	3	9	0.00521	23.811 mins
RSP exact 4 q	49.241 secs	3	9	0.01759	33.099 mins
WOP 2 q	0.452 secs	2	6	0.03491	24.324 mins
WOP 3 q	0.474 secs	3	9	0.00519	23.811 mins
WOP 4 q	0.638 secs	3	9	0.01728	33.099 mins
OP 2 q	0.239 secs	2	6	0.00481	24.324 mins
OP 3 q	0.249 secs	3	9	0.00518	23.811 mins
OP 4 q	0.344 secs	3	9	0.01727	33.099 mins
MatchAlign DL 2 q	0.693 secs	2	6	0.01171	3.941 mins
MatchAlign DL 3 q	0.547 secs	3	9	0.00703	3.762 mins
MatchAlign DL 3 q	0.534 secs	3	9	0.01036	11.065 mins
MatchAlign MGSP	4.640 secs	26.2	9	0.01634	35.536 mins
BEFA with Huang and Wand prior	3.346 secs	3	9	0.00254	18.875 mins
BEFA with pre-specified S	3.448 secs	3	9	0.00253	18.836 mins

As it can be seen from table 3.1, Orthogonal Procrustes is the method requiring on average the least time in order to solve the rotational ambiguities and the label switching problem. All the models not selecting the number of factors except the models with pre-specified number of factors equal to 2 are able to correctly identify the true number of factors. Regarding the models selecting the number of factors, BEFA with Huang and Wand Prior for the covariance matrix and BEFA with pre-specified values of S for the

Inverse-Wishart prior of the covariance matrix identify the correct number of factors. Additionally, the Multiplicative Gamma Process Shrinkage model with MatchAlign method overestimates the number of factors. However, the estimated number of active factors coincides with the true number of factors. All the models except the ones with pre-specified number of factors equal to 2 are able to correctly load the observed variables. The method with the smallest on average normed dissimilarity between the posterior mean of the covariance matrix and the covariance matrix estimated by the posterior mean of the factor loadings matrix Λ after solving rotational ambiguities and label switching problems is BEFA.

Regarding diagram 3.1, each column depicts one factor and each row one observed variables. Moreover, the colouring of each cell indicates whether the factor loading is negative (Blue) or positive (Green) and the brightness indicates the degree of negativity or positivity. The 3.1 diagram shows that the Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the factors equal to $5 \log 9$ overestimates the number of factors, but the estimated number of active factors is equal to the true number of factors and loads the observed variables correctly. In particular, only three factors have measurements loaded onto them indicating that the number of active factors is equal to 3. Furthermore, it is shown that observed variables 1,2 and 3 are loaded to one factor, 4,5 and 6 to another factor and 7,8 and 9 to a third factor.

Regarding the following diagrams 3.2,3.3 and 3.4 each window represents one factor and every black dot in each window represents the ergodic mean of the factor loading of that specific factor. Furthermore, the blue intervals are the 99% Higher Posterior Density intervals and the red intervals are the simultaneous 99% credible regions. From the following 3.2,3.3 and 3.4 diagrams we can acquire the following information: the Normal inverse-Gamma model with the RSP,WOP and OP methods select the correct number of factors and load the observed variables correctly. Specifically, from the diagrams 3.2(a),3.3(a) and 3.4(a) it is noticed that some variables are not loaded onto any factor, indicating that more than 2 factors are needed. Furthermore, from the diagrams 3.2(c),3.3(c) and 3.4(c) it is perceived that for each model when the number of factors is equal to 4, there is a redundant factor. A redundant factor is a factor without any observed variable loaded at it. Conclusively, as mentioned in the first chapter of this thesis regarding the selection of the number of factors, we can identify that the correct number of factors is equal to 3. Additionally, we can perceive from the following diagrams 3.2(b),3.3(b),3.4(b)3.2(c),3.3(c) and 3.4(c) that the models

load the observed variables correctly. In order to make it more perceptible, an example is illustrated with the same logic also applying to the rest of the diagrams. For example, in the diagram 3.2(b) we can see that in the first window the 7,8 and 9 observed variables are loaded correctly to the same factor, in the second window the 4,5 and 6 observed variables are loaded correctly to the same factor and in the third window the 1,2 and 3 observed variables are loaded correctly to the same factor.

Concerning the 3.5 diagram, each column represents one factor and each row one observed variable. Additionally, the colouring of each box indicates whether the factor loading is negative (Blue) or positive (Green) and the brightness indicates the degree regarding the negativity or positivity. From the 3.5 diagram we can acquire the following information: the Dirichlet Laplace model with MatchAlign method select the correct number of factors and load the observed variables correctly. In particular, from the 3.5(a) diagram it is seen that some variables are not loaded onto any factor indicating that more than 2 factors are needed. Additionally, from the diagram 3.5(c) we can perceive that there is a redundant factor. Conclusively, as mentioned in the first chapter of this thesis regarding the selection of the number of factors, we can identify that the correct number of factors is equal to 3. Moreover, we can perceive from the following diagrams 3.5(b) and 3.5(c) that the model allocates the observed variables correctly. Specifically, it is seen from 3.5(b) or 3.5(c) that the 1,2,3 observed variables are in the same column indicating their allocation to the same factor. Similarly, 4,5,6 observed variables and 7,8,9 observed variables appear in different columns indicating their allocation to same factors respectively.

Figure 3.1: Multiplicative Gamma Process Shrinkage model with MatchAlign identification method and initial value for the factors equal to $5 \log_{10} 9$ for scenario 1

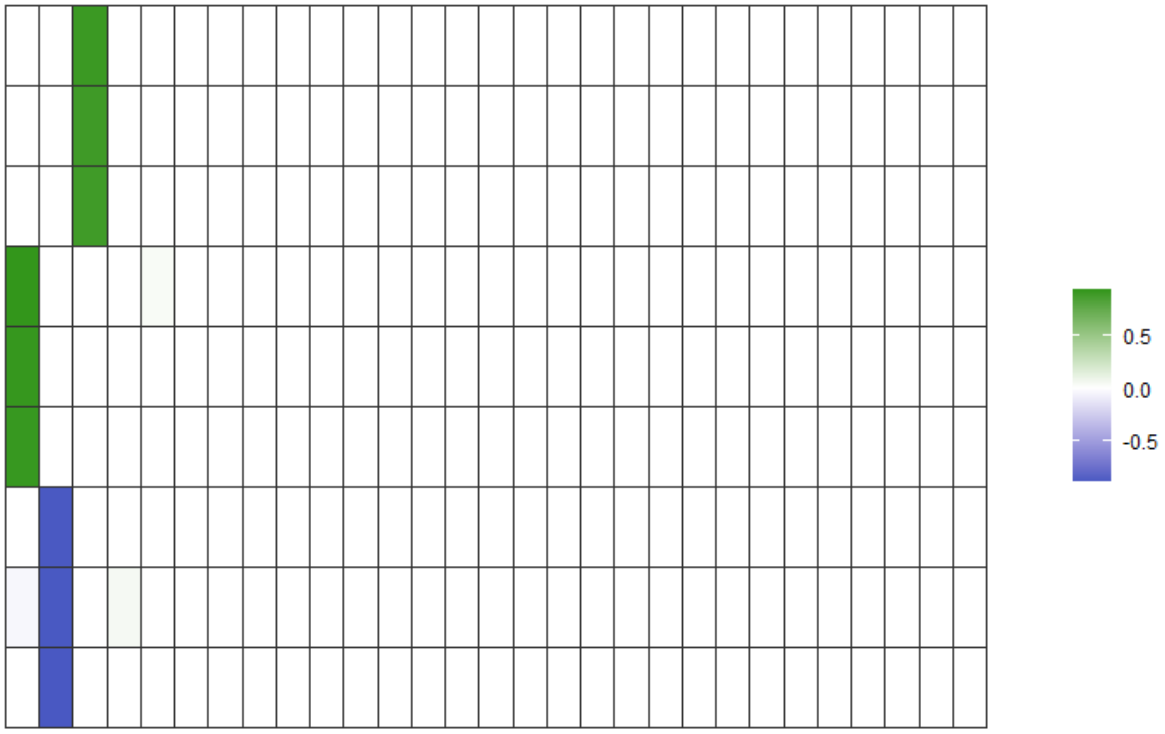
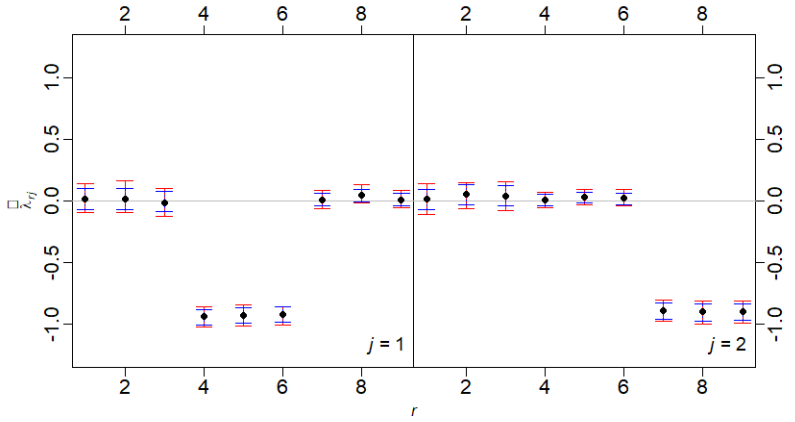
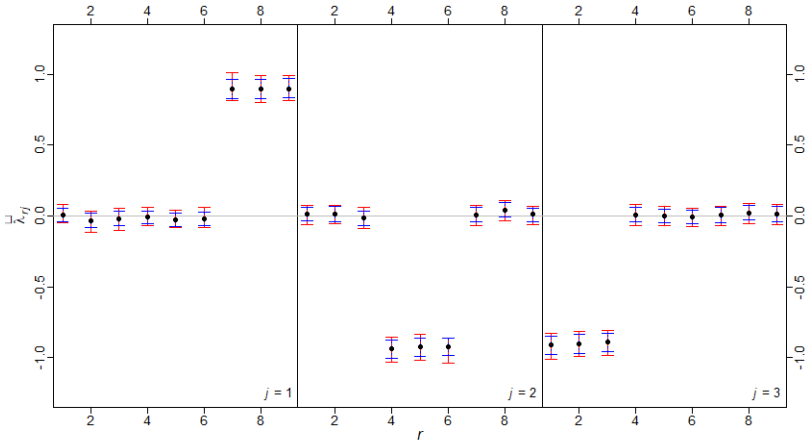


Figure 3.2: RSP With 2,3 and 4 factors for scenario 1

(a) RSP with 2 factors



(b) RSP with 3 factors



(c) RSP with 4 factors

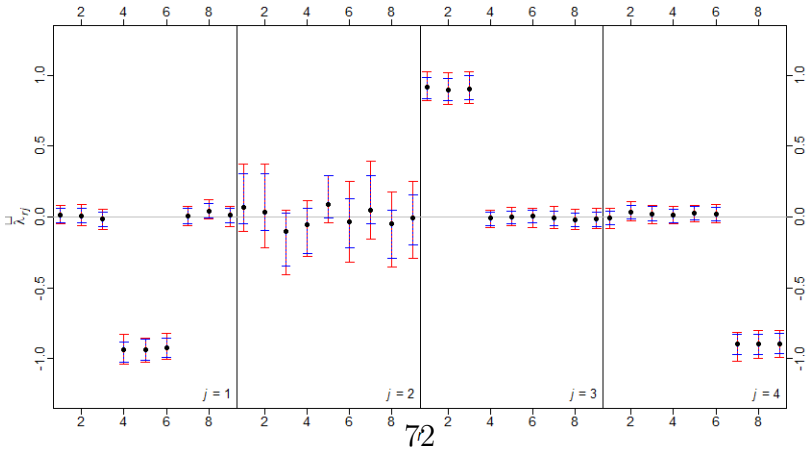
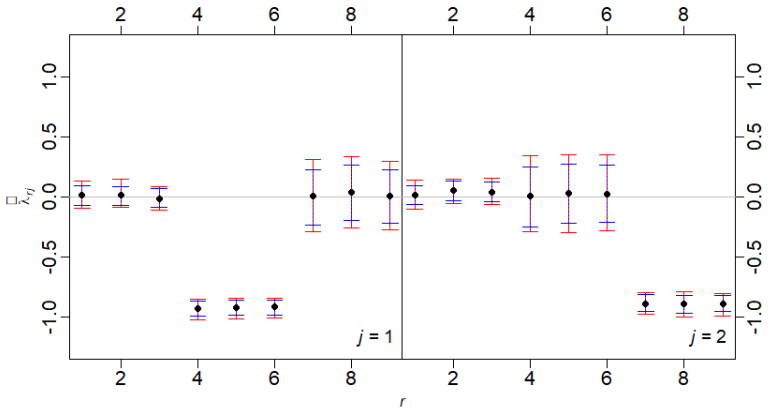
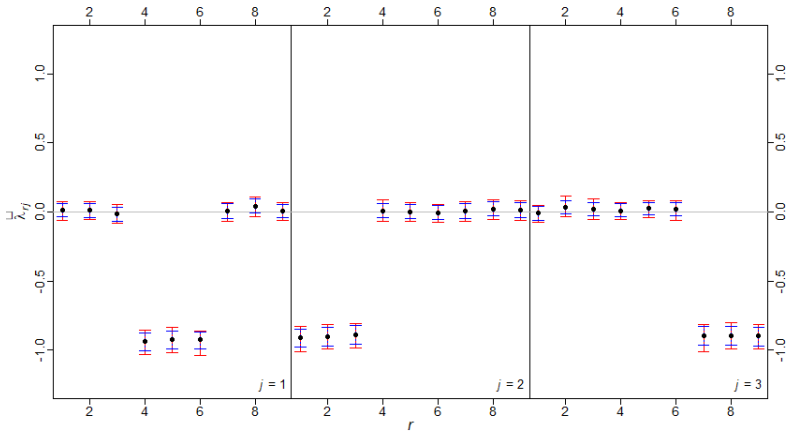


Figure 3.3: Weighted Orthogonal Procrustes With 2,3,4 factors for scenario 1

(a) WOP with 2 factors



(b) WOP with 3 factors



(c) WOP with 4 factors

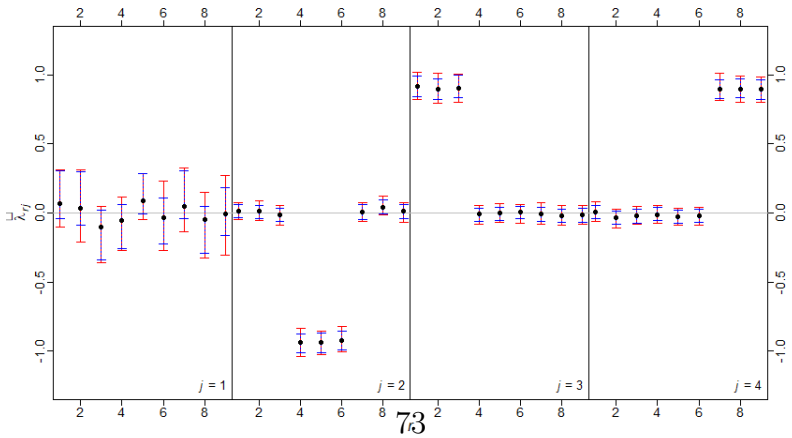
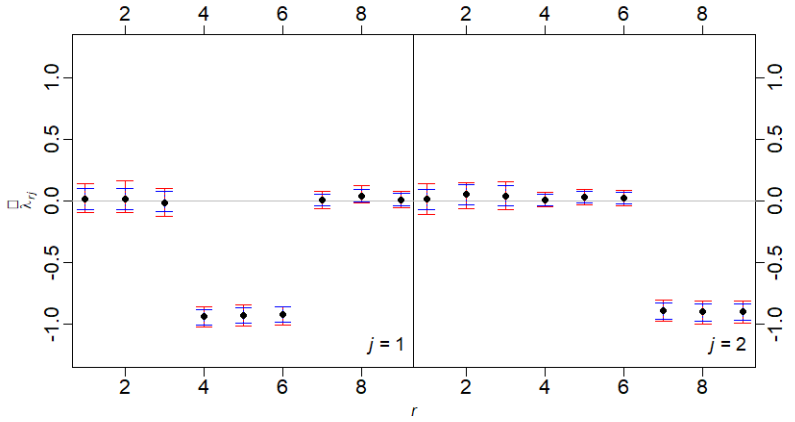
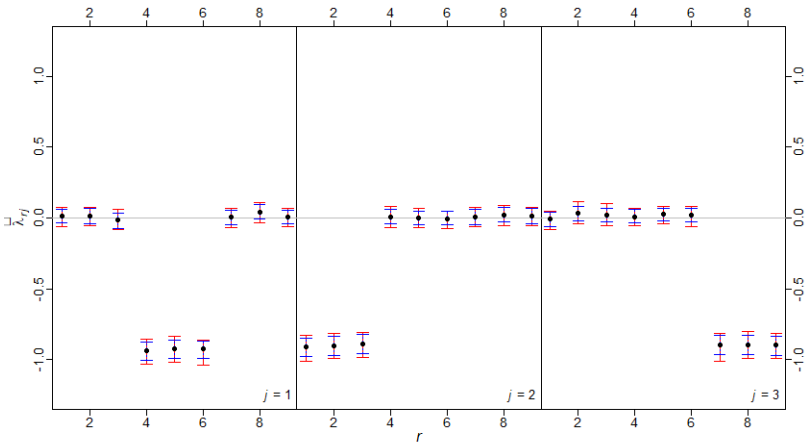


Figure 3.4: Orthogonal Procrustes With 2,3,4 factors for scenario 1

(a) OP with 2 factors



(b) OP with 3 factors



(c) OP with 4 factors

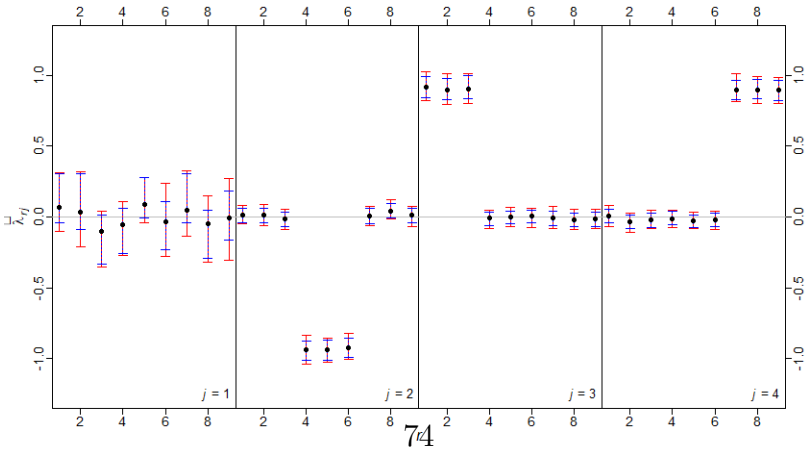
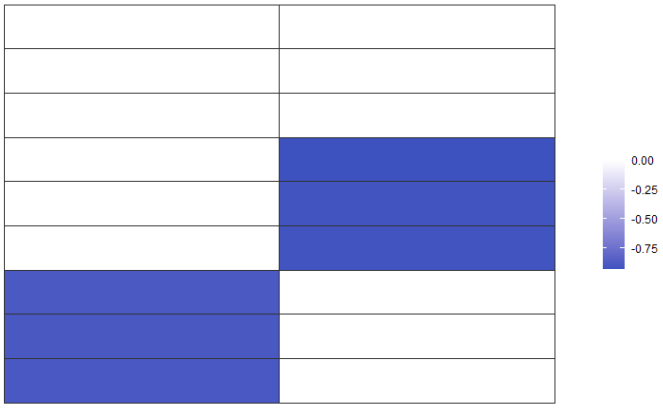


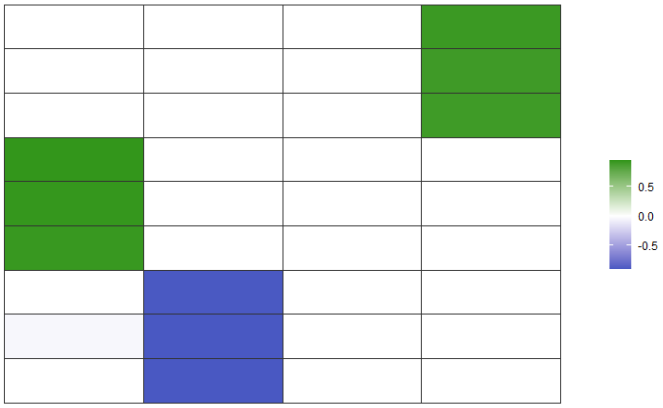
Figure 3.5: Dirichlet Laplace model with MatchAlign identification method and number of factors 2,3 and 4 for scenario 1

(a) MatchAlight with 2 factors



(b) MatchAlight with 3 factors

(c) MatchAlight with 4 factors



3.2 Synthetic data scenario 2

Table 3.2: Summarized results for synthetic data scenario 2

Model and identification method	Time to solve identification issues	Factors by model	Variables loaded correctly	Metric	Time to simulate MCMC samples
PLT with 8 factors	-	8	24	-	2.384 hours
PLT with 9 factors	-	9	27	-	1.943 hours
PLT with 10 factors	-	9	27	-	1.997 hours
RSP FSA 8 q (sa=1000)	5.580 mins	8	24	0.02831	1.546 hours
RSP FSA 9 q (sa=1000)	6.418 mins	9	27	0.02731	1.753 hours
RSP FSA 10 q (sa=1000)	6.077 mins	9	27	0.06098	1.877 hours
RSP PSA 8 q (sa=250)	42.366 mins	8	24	0.02781	1.546 hours
RSP PSA 9 q (sa=250)	42.198 mins	9	27	0.02672	1.753 hours
RSP PSA 10 q (sa=250)	46.377 mins	9	27	0.06099	1.877 hours
WOP 8 q	1.314 secs	8	24	0.05761	1.546 hours
WOP 9 q	1.331 secs	9	27	0.02634	1.753 hours
WOP 10 q	1.975 secs	9	27	0.03637	1.877 hours
OP 8 q	1.282 secs	8	24	0.05760	1.546 hours
OP 9 q	0.476 secs	9	27	0.02613	1.753 hours
OP 10 q	1.963 secs	9	27	0.03637	1.877 hours
MatchAlign DL 8 q	1.069 secs	8	24	0.03006	19.133 mins
MatchAlign DL 9 q	1.095 secs	9	27	0.02642	15.476 mins
MatchAlign DL 10 q	1.714 secs	9	27	0.03766	24.447 mins
MatchAlign MGSP	4.358 secs	27.8	27	0.01939	1.3 hours
BEFA with Huang and Wand prior	11.339 secs	9	27	0.00471	1.511 hours
BEFA with pre-specified S	11.649 secs	9	27	0.00472	1.645 hours

As it can be seen from table 3.2, Orthogonal Procrustes is the method requiring on average the least time in order to solve the rotational ambiguities and the label switching problem. All the models not selecting the number of factors except the ones with pre-specified number of factors equal to 8 are able to correctly identify the true number of factors. Regarding the models selecting the number of factors, BEFA with Huang and Wand Prior for the covariance matrix and BEFA with pre-specified values of S for the Inverse-Wishart prior of the covariance matrix identify the correct number

of factors. Furthermore, the Multiplicative Gamma Process Shrinkage model with MatchAlign method overestimates the number of factors. However, the estimated number of active factors coincides with the true number of factors. All the models except ones with pre-specified number of factors equal to 8 are able to load correctly the observed variables. The method with the smallest on average normed dissimilarity between the posterior mean of the covariance matrix and the covariance matrix estimated by the posterior mean of the factor loadings matrix Λ after solving rotational ambiguities and label switching problems is BEFA with the Huang and Wand prior on S for the Inverse-Wishart prior of the covariance matrix.

Diagram 3.6 shows that the Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the factors equal to $5 \log 27$ overestimates the number of factors, but the estimated number of active factors is equal to the true number of factors and allocates the observed variables correctly. The explanation of this conclusion is analogous to the one given in the previous subsection of this chapter(3.1).

As it can be inspected from the diagrams 3.7, 3.8, A.1 and A.2 the Normal inverse-Gamma model with Full Simulation Annealing RSP, Partial Simulation Annealing RSP, Weight Orthogonal Procrustes and Orthogonal Procrustes methods identify the correct number of factors and allocate the observed variables correctly. The explanation of this conclusion is analogous to the explanation given in the previous subsection of this chapter(3.1). As mentioned in the second chapter, diagrams 3.7 and 3.8 indicate that the solution of the Partial Simulation Annealing RSP method is more efficient than these of the Full Simulation Annealing RSP. This can be further supported by inspecting the 3.8 and 3.7 diagrams and observing that the 99% Higher Posterior Density intervals and the simultaneous 99% credible regions are both bigger in the Full Simulation Annealing RSP method indicating more uncertainty.

Diagram 3.9 shows that the Dirichlet Laplace model with MatchAlign method selects the correct number of factors and allocates the observed variables correctly. The explanation of this conclusion is analogous to the one given in chapter(3.1).

Figure 3.6: Multiplicative Gamma Process Shrinkage model with MatchAlign identification method and initial value for the factors equal to $5 \log_{10} 27$ for scenario 2

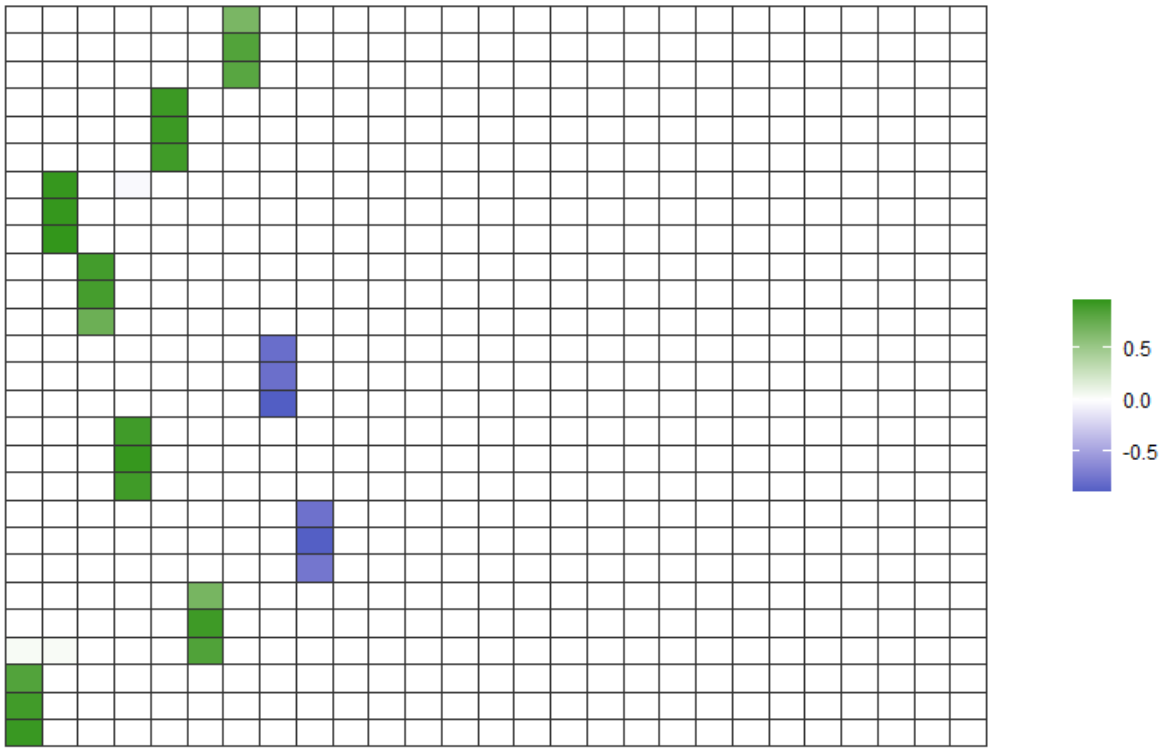
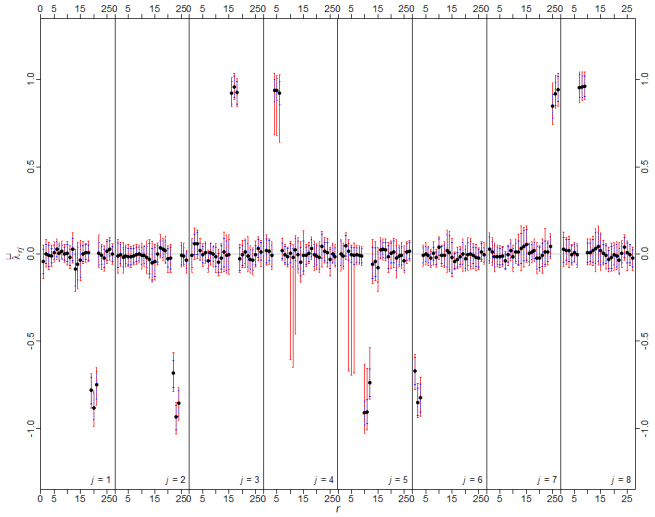
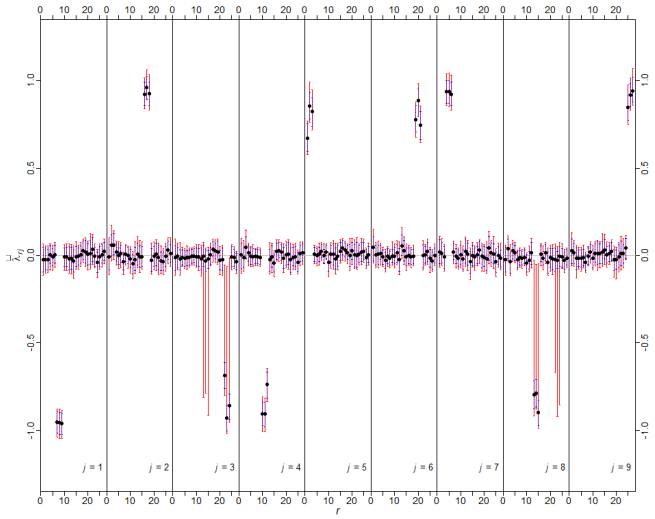


Figure 3.7: Full Simulation Annealing With 8,9 and 10 factors for scenario 2

(a) Full Simulation Annealing RSP with 8 factors



(b) Full Simulation Annealing RSP with 9 factors



(c) Full Simulation Annealing RSP with 10 factors

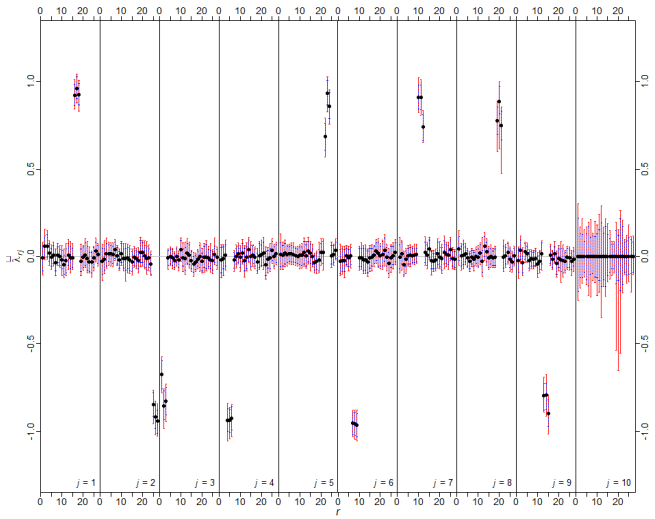
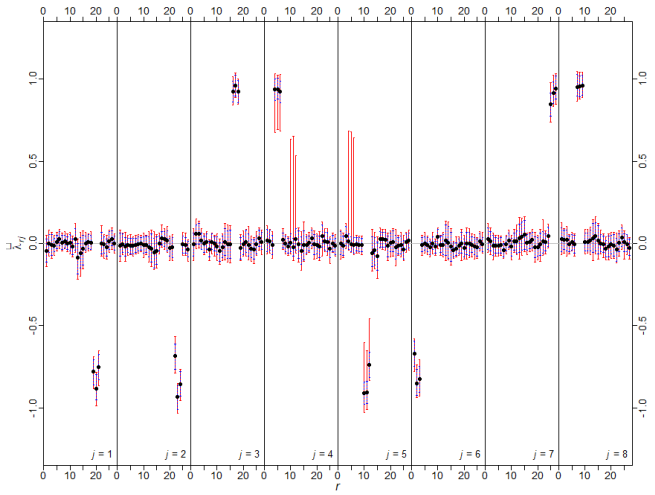
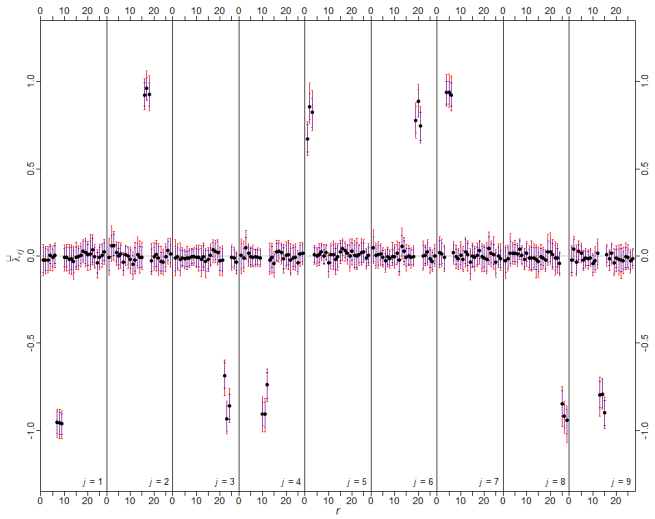


Figure 3.8: Partial Simulation Annealing With 8,9 and 10 factors for scenario 2

(a) Partial Simulation Annealing RSP with 8 factors



(b) Partial Simulation Annealing RSP with 9 factors



(c) Partial Simulation Annealing RSP with 10 factors

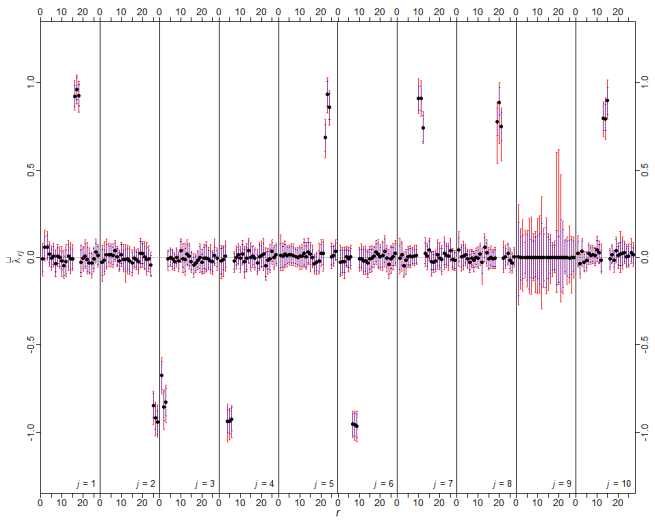
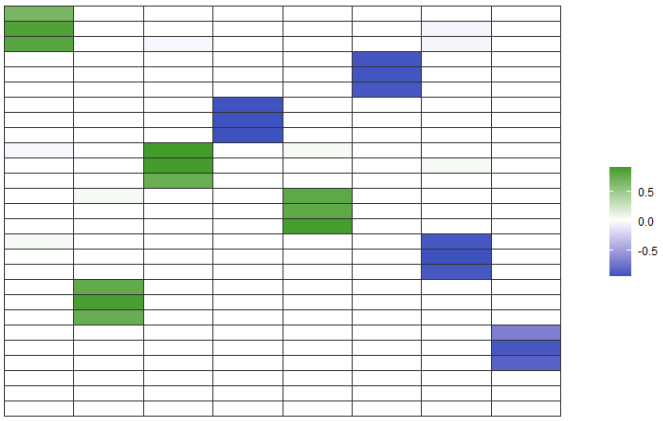
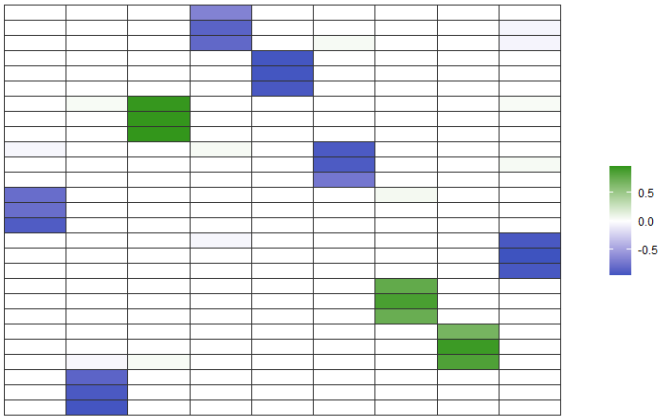


Figure 3.9: Dirichlet Laplace model with MatchAlign identification method and number of factors 8,9 and 10 for scenario 2

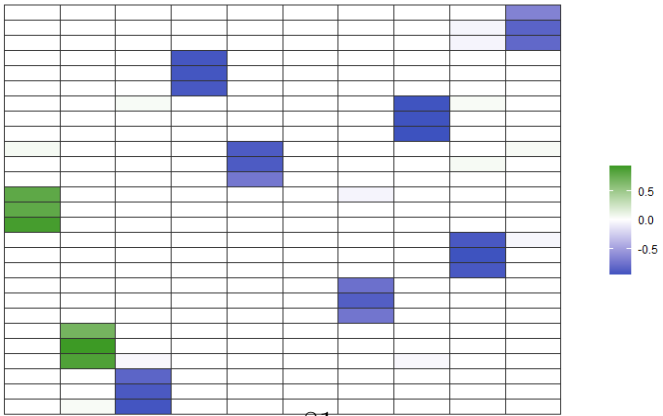
(a) MatchAlight with 8 factors



(b) MatchAlight with 9 factors



(c) MatchAlight with 10 factors



3.3 Synthetic Data Scenario 3

Table 3.3: Summarized results for synthetic data scenario 3

Model and identification method	Time to solve identification issues	Factors by model	Variables loaded correctly	Metric	Time to simulate MCMC samples
PLT 19 q	-	19	114	-	20.565 hours
PLT 20 q	-	20	120	-	22.522 hours
PLT 21 q	-	20	120	-	1.027 days
RSP FSA 19 q (sa=1000)	1.626 hours	19	114	2.72681	20.706 hours
RSP FSA 20 q (sa=1000)	1.004 hours	20	120	1.67841	21.805 hours
RSP FSA 21 q (sa=1000)	1.372 hours	20	120	1.07482	1.021 days
RSP PSA 19 q (sa=250)	1.724 hours	19	114	0.46759	20.706 hours
RSP PSA 20 q (sa=250)	2.403 hours	20	120	0.47897	21.805 hours
RSP PSA 21 q (sa=250)	1.912 hours	20	120	0.49941	1.021 days
WOP 19 q	11.725 secs	19	114	0.46009	20.706 hours
WOP 20 q	11.805 secs	20	120	0.45894	21.805 hours
WOP 21 q	12.337 secs	20	120	0.48243	1.021 days
OP 19 q	3.928 secs	19	114	0.44282	20.706 hours
OP 20 q	3.691 secs	20	120	0.45434	21.805 hours
OP 21 q	4.208 secs	20	120	0.47807	1.021 days
MatchAlign DL 19 q	12.131 secs	19	114	0.41566	1.466 hours
MatchAlign DL 20 q	6.648 secs	20	120	0.42566	1.648 hours
MatchAlign DL 21 q	12.625 secs	20	120	0.44887	1.954 hours
MatchAlign MGSP	7.127 secs	34	120	0.11293	4.363 hours
BEFA with Huang and Wand prior	2.412 mins	20	120	11.1919	1.011 days
BEFA with pre-specified S	1.328 mins	20	120	11.17475	23.806 hours

As it can be seen from table 3.3, Orthogonal Procrustes is the the method requiring on average the least time in order to solve the rotational ambiguities and the label switching problem. All the models not selecting the number of factors except those with pre-specified number of factors equal to 19 are able to correctly identify the true number of factors. Regarding the models selecting the number of factors, BEFA with Huang and Wand Prior for the covariance matrix and BEFA with pre-specified values of S for the Inverse-Wishart prior of the covariance matrix identify the correct number

of factors. Furthermore, the Multiplicative Gamma Process Shrinkage model with MatchAlign method overestimates the number of factors. However, the estimated number of active factors coincides with the true number of factors. All the models except those with pre-specified number of factors equal to 19 are able to correctly allocate the observed variables. The method with the smallest on average normed dissimilarity between the posterior mean of the covariance matrix and the covariance matrix estimated by the posterior mean of the factor loadings matrix Λ after solving rotational ambiguities and label switching problems is MatchAlign with the Multiplicative Gamma Shrinkage prior.

Diagram 3.10, shows that that the Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the factors equal to $5 \log 120$ overestimates the number of factors, but the estimated number of active factors is equal to the true number of factors and allocates the observed variables correctly. The explanation of this conclusion is analogous to the one given in a previous subsection of this chapter(3.1).

Diagram 3.11 shows that Dirichlet-Laplace model with MatchAlign method selects the correct number of factors and allocates the observed variables correctly. The explanation of this conclusion is analogous to the one given in chapter(3.1)

Figure 3.10: Multiplicative Gamma Process Shrinkage model with MatchAlign identification method and initial value for the factors equal to $5 \log_{10} 120$ for scenario 3

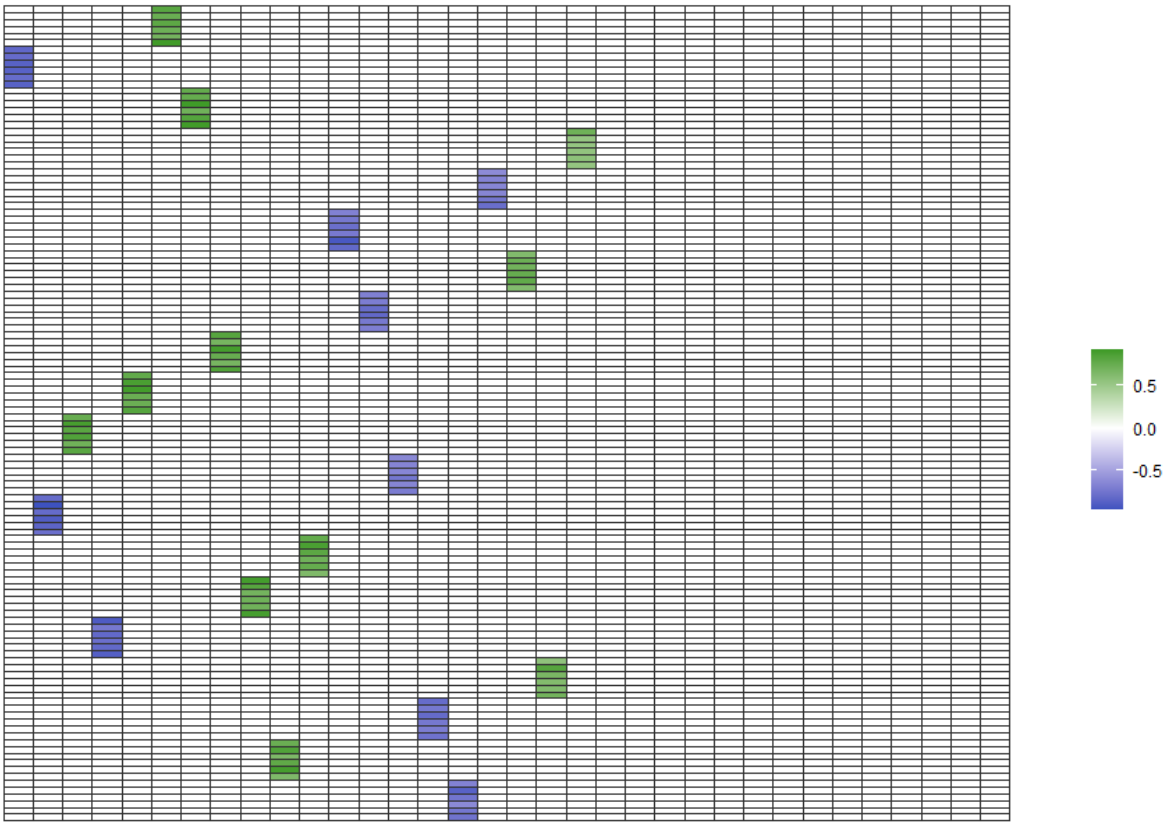
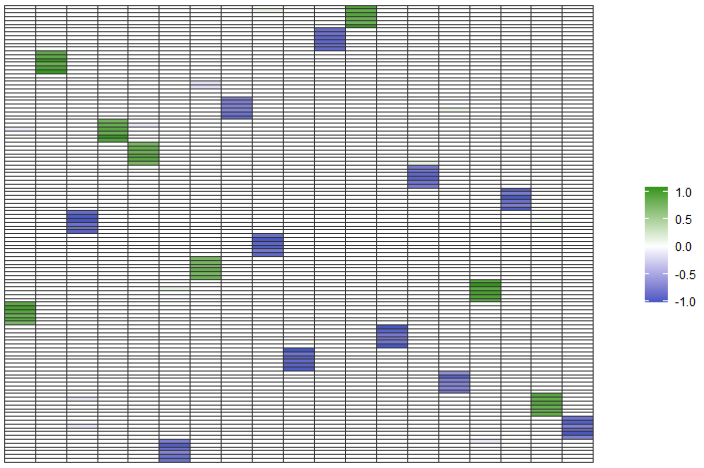
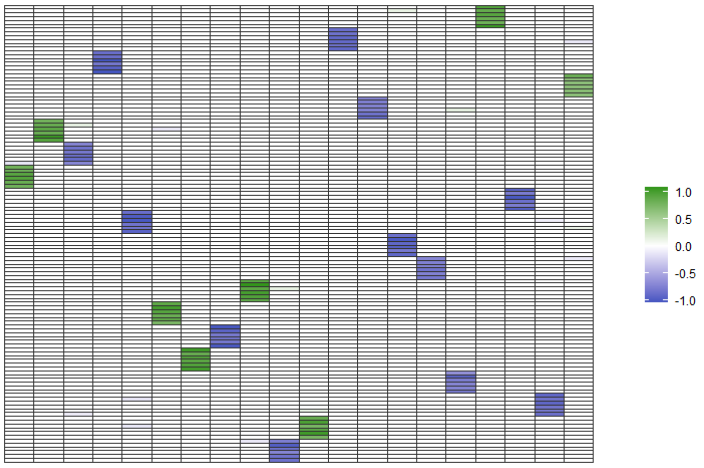


Figure 3.11: Dirichlet Laplace model with MatchAlign identification method and number of factors 19,20 and 21 for scenario 3

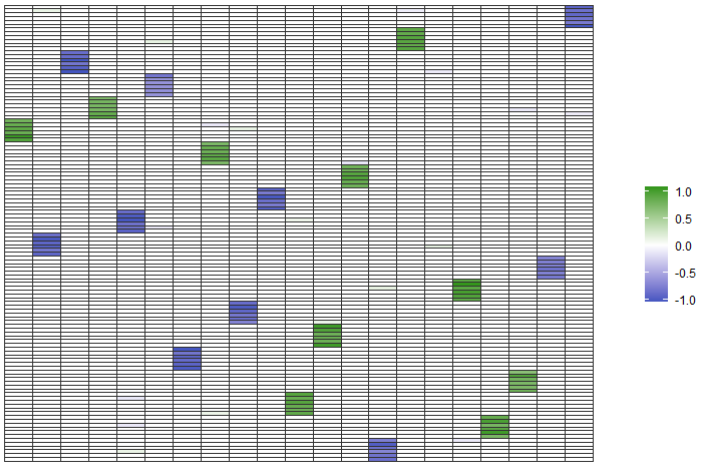
(a) MatchAlight with 19 factors



(b) MatchAlight with 20 factors



(c) MatchAlight with 21 factors



3.4 Final thoughts on the results of the three scenarios

From the three synthetic scenarios the following conclusions are elicited. The Dirichlet-Laplace model with MatchAlign method is the one simulating the MCMC samples faster among the models not selecting the number of factors. Regarding the models selecting the number of factors, it can be seen that for small numbers of observed variables, the BEFA models produce the MCMC samples faster, but for medium and large numbers of observed variables, the Multiplicative Gamma Process Shrinkage model with MatchAlign method is the better option. Regarding the method resolving the rotational ambiguities and the label switching problem faster in all three scenarios, the Orthogonal Procrustes method achieves it. Regarding the selection of the factors in all three scenarios, the models not selecting the number of factors with predetermined number of factors equal to or greater than the true number of factors are able to identify the correct number of factors. Regarding the models selecting the number of factors, BEFA with Huang and Wand Prior for the covariance matrix and BEFA with pre-specified values of S for the Inverse-Wishart prior of the covariance matrix identify the correct number of factors in all three scenarios. However, the Multiplicative Gamma Process Shrinkage model with MatchAlign method overestimates the number of factors in all three scenarios, but the number of active factors were equal to the true number of factors. All the models except the ones with pre-specified number of factors less than the true number of factor allocate the observed variables correctly in all three scenarios. All the models construct Chains reaching convergences. Regarding the metric for small and medium numbers of observed variables, the BEFA methods produce more sufficient solutions, but for large numbers of observed variables the MatchAlign method produces the most sufficient solutions. Nevertheless, the previous mentioned conclusions are based on ideal scenarios where the observed variables are only high correlated if they belong to the same factor and low correlated otherwise. On a real data set the findings may differ from the findings presented in this chapter. For this reason, in the next chapter the models will be compared with each other upon real datasets.

Chapter 4

Comparison of the different models in real data sets

The purpose of this chapter is the same as Chapter 3, but, instead of synthetic datasets, real ones will be employed. The comparison will be made in the same fields. For ease of presentation, the fields will be reintroduced as follows: 1) How much time each method requires in order to solve rotational ambiguities and label switching. 2) The number of factors selected by the models in comparison with the true number of factors according to R. A. Martin et al. (2003) for the Humor Style Questionnaire and to Goldberg (1992) for the Big Five Personality Test. 3) the number of variables which are allocated correctly according to R. A. Martin et al. (2003) for the Humor Style Questionnaire and to Goldberg (1992) for the Big Five Personality Test. 4) A metric evaluating the performance of solving rotational ambiguities and label switching.

All coding was conducted on the R programming language (Team, 2021). As already mentioned for the generation of the MCMC samples regarding the following methods : (RSP Exact,RSP Full Simulation Annealing, RSP Partial Simulation Annealing,WOP,OP) the MCMCpack ("MCMCpack", n.d.) package of R will be used with number of iterations equal to 1000000, burn in equal to 400000 and thinning equal to 150. For the generation of the MCMC samples regarding the PLT method the MCMCpack ("MCMCpack", n.d.) package of R will be employed with number of iterations equal to 1000000, burn in equal to 400000 and thinning equal to 150. For the generation of the MCMC samples for the MatchAlign method with the priors (Dirichlet-Laplace Shrinkage Prior, Multiplicative Gamma Shrinkage Prior) specified in the second chapter of this thesis the infinitefactor (Poworoznek, 2020) package of R will be used with number of iterations equal to 1000000, burn

in equal to 400000,thinning equal to 150 and initial value for the number of factor for the Multiplicative Gamma Shrinkage Prior equal to $5 \log_{10} p$ as suggested by the authors in Bhattacharya and Dunson (2011). For the generation of the MCMC samples for the BEFA method the BayesFM (Piatek, 2021) package of R will be used with number of iterations equal to 1000000 and burn in equal to 800000. Furthermore, for the methods which do not choose the number of factors, only three chains will be presented having the number of factors equal to $(q-1, q, q+1)$ where q is the true number of factors. The factor.switching (Papastamoulis & Ntzoufras, 2022b) packages of R which will be used for the implementation of the solution of the rotational ambiguities and label switching by these methods: RSP Exact, RSP Full Simulation Annealing, RSP Partial Simulation Annealing, WOP, OP. Similarly, the infinitefactor (Poworoznek, 2020) package of R will be used for MatchAlign method. Moreover, the BayesFM (Piatek, 2021) package of R will be used for BEFA method. The PLT method does not require a method to solve rotational ambiguities and label switching because of the restrictions implemented in the generation of the MCMC samples.

The prior specifications for the models producing the posterior samples is the following

The prior distributions for the generation of the posterior samples by the Gibbs sampler of the MCMCpack are

For the rows of the factor loadings matrix Λ

$$\Lambda_{ij} \sim N(l_{0_{ij}}, {L_{0_{i,j}}}^{-}1) \text{ for } i = 1, \ldots, p,\ j = 1, \ldots, q$$

with value for $L^{-}_{0_{i,j}}1$ equal to 0 and for $l_{0_{ij}}$ also equal to 0. By setting $L^{-}_{0_{i,j}}1 = 0$ we are employing an improper prior.

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_i \sim IG(\frac{a_0}{2}, \frac{b_0}{2}),\ i = 1, \ldots, p$$

with value for a_0 and b_0 equal to 0.001.

The prior distributions for the generation of the posterior samples by the Positive Lower Triangular method are the following

For the non-diagonal elements of the factor loadings matrix

$$\Lambda_{ij} \sim N(0, \infty) \; \forall i \neq j \; i = 1, \ldots, p, \; j = 1, \ldots, q.$$

By setting the variance of the non-diagonal elements of the factor loadings matrix to ∞ we are employing an improper prior.

For the diagonal elements of the factor loadings matrix

$$\Lambda_{jj} \sim N(0, \infty) 1(\Lambda_{ii} > 0) \; j = 1, \ldots, q.$$

This prior specification is a truncation of the normal distribution restricting the distribution to only positive values. By setting the variance of the diagonal elements of the factor loadings matrix to ∞ we are employing an improper prior.

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_i^2 \sim IG(\frac{0.001}{2}, \frac{0.001}{2}), \; i = 1, \ldots, p.$$

The prior distributions for the generation of the posterior samples by the MatchAlign method with Dirichlet-Laplace prior are the following

For the rows of the factor loadings matrix Λ

$$\lambda_{ik} | \varphi_{ik}, t_i \sim DE(\varphi_{ik} t_i) \; k = 1, \ldots, K$$

$$\varphi_i \sim Dir(1/2, \ldots, 1/2) \; t_i \sim G(K\frac{1}{2}, \frac{1}{2})$$

where $i = 1, \ldots, p$,$\varphi_i = (\varphi_{i1}, \ldots, \varphi_{iK})$ and K is the maximum number of factors.

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_i \sim IG(\frac{1}{2}, \frac{1}{2}), \; i = 1, \ldots, p.$$

The prior distributions for the generation of the posterior samples by the MatchAlign method with Multiplicative Gamma Process Shrinkage Prior are

the following

For the factor loadings matrix

$$\lambda_{jh}|\Phi_{jh}, t_h \sim N(0, \Phi_{jh}^{-1} t_h^{-1}) \ , \ \Phi_{jh} \sim Gamma(\frac{3}{2}, \frac{3}{2}) \ , \ t_h = \prod_{l=1}^{h} \delta_l$$

$$\delta_1 \sim Gamma(1,1) \ , \ \delta_l \sim Gamma(2,1) \ l \geq 2 \ , \ j = 1, \ldots, p, l = 1, \ldots, h.$$

For the diagonal elements of the variance matrix Σ of the idiosyncratic errors

$$\sigma_j^{-2} \sim Gamma(1, 0.3) \ , \ j = 1, \ldots, p.$$

The prior distributions for the generation of the posterior samples by BEFA method are the following

The prior of the binary indicator matrix

$$Pr(\Delta_m = e_l | t_l) = t_l \ , \ \sum_{l=0}^{q} t_l = 1 \ , \ l = 1, 2, \ldots, q$$

where Δ_m , for $m = 1, 2, \ldots p$ is the m-th row of the binary indicator matrix Δ and e_k is the indicator vector

$$t^\star \sim Dir(\frac{1}{K}, \frac{1}{K} \ldots, \frac{1}{K}) \ , \ t_m = (t_{0m}, (1 - t_{0m}) t_1^\star, (1 - t_{0m}) t_2^\star, \ldots, (1 - t_{0m}) t_q^\star)$$

$$\text{for } m = 1, 2, \ldots, p \text{ and } t_{0m} \sim Beta(2, 1)$$

where K is the maximum number of factors.

The prior distribution of the diagonal elements of the variance matrix Σ of the idiosyncratic errors is

$$\sigma_m^2 \sim IG\,(2, 1) \ \ m = 1, 2, \ldots, p.$$

The prior distribution of the only non-zero factor loadings in the m-th row is

$$\Lambda_{ml}^{\Delta} | \sigma_m^2 \sim N(0, 10\sigma_m^2), m = 1, \ldots, p.$$

The prior distribution of the covariance matrix Z with pre-specified S for the Inverse-Wishart prior is

$$Z \sim IW(K+1, diag(1))$$

where K is the maximum number of factors and $diag(1)$ is a $q \times q$ diagonal matrix with all objects equal to 1.

The prior distribution of the covariance matrix Z with the Huang and Wand prior in S for the Inverse-Wishart prior is

$$Z \sim IW(K+1, W)\ W = diag(w_1, w_2, \ldots, w_q)$$

$$\text{where } w_k \sim G(\frac{1}{2}, \frac{1}{4})$$

where K is the maximum number of factors.

Finally, the intermediate steps S are equal to

$$S = 1 + \Phi$$

$$\text{where } \Phi \sim Poisson(4).$$

4.1 Humor Style Questionnaire

R. A. Martin (2001) defined as sense of humor everything an individual performs or vocalizes which is understood as humorous and leads other individuals to laugh, implicating those mental processes aiming at producing and understanding funny stimulations and their effective reactions. The Humor Style questionnaire is a self-reporting inventory quantifying individual dissimilarities for each one of the four humor types represented in the humor style model proposed by R. A. Martin et al. (2003). The original questionnaire has 32 questions and the responses are in a seven point Likert scale varying from 1 (totally disagree) to 7 (totally agree). However, a five point Likert scale varying from 1 (Never or very rarely true) to 5 (Very often or always true) having the same questions as the original will be used. Each of the four humor types is linked to 8 questions. Given that the phrases in the questions are large, the question index will be used in order to describe the links between the questions and the humor types. Specifically, affiliative humor is linked to questions (1,5,9,13,17,21,25,29), self-enhancing humor to (2,6,10,14,18,22,26,30), aggressive humor to (3,7,11,15,19,23,27,31) and self-defeating to (4,8,12,16,20,24,28,32). Before proceeding, a brief description of humor styles will be made. Affiliative humor is relevant to the employment of humor in order to minimize disputes, form stronger bonds and make

other individuals happy. Self-enhancing humor is affiliated with the use of humor in order to detach yourself from stressful events and cope with stressors positively. Aggressive humor is affiliated with the use of humor aiming at deriding other individuals and making them seem ridiculous. Finally, self-defeating humor is affiliated with the use of humor in order to humiliate yourself or accept humiliation in order to be accepted by your peers and friends. R. A. Martin et al. (2003) considered affiliative and self-enhancing humor as an indication of psychological well-being and aggresive and self-defeating humor as harmful. Furthermore, self-enhancing and self-defeting humor are intrapersonal while affiliative and aggresive humor are interpersonal. Consequently, the humor style questionnaire data set should have four latent factors, each of which representing one particular humor style.

The data set employed for our analysis can be acquired from `https://openpsychometrics.org/`. The collection of the data has been made from an interactive online version of the Humor Style Questionnaire. The questions of the Humor Style Questionnaire can be found in the appendix A.1. Apart from the questionnaire, the respondents had to state their age and choose their gender from a drop down list, the available options being 1(male), 2(female) and 3(other). The data set consists of 1071 respondents: 581 males, 477 females and 8 other. Furthermore, the age of the respondents ranges from 14 to 70. Prior to analysing the data, respondents who not having fully completed the questionnaire were removed and the data was standardised. Furthermore, false entries were checked for. The remaining respondents are 993: 537 males, 443 females and 8 other, with their age ranging from 14 to 70

Table 4.1: Summarized results for Humor Style Questionnaire

Model and identification method	Time to solve identification issues	Factors by model	Variables loaded correctly	Metric	Time to simulate MCMC samples
PLT 3 q	-	3	23	-	54.069 mins
PLT 4 q	-	4	31	-	1.041 hours
PLT 5 q	-	4	31	-	1.455 hours
RSP exact 3 q	32.141 secs	3	24	0.01997	52.954 mins
RSP exact 4 q	1.420 mins	4	32	0.02296	1.007 hours
RSP exact 5 q	3.706 mins	4	32	0.41135	1.613 hours
WOP 3 q	1.973 secs	3	24	0.01993	52.954 mins
WOP 4 q	2.077 secs	4	32	0.02081	1.007 hours
WOP 5 q	3.087 secs	4	32	0.20717	1.613 hours
OP 3 q	0.547 secs	3	24	0.01788	52.954 mins
OP 4 q	0.595 secs	4	32	0.02062	1.007 hours
OP 5 q	0.857 secs	4	32	0.20582	1.613 hours
MatchAlign DL 3 q	1.286 secs	3	24	0.01824	16.914 mins
MatchAlign DL 4 q	1.417 secs	4	32	0.02153	9.024 mins
MatchAlign DL 5 q	3.426 secs	4	32	0.26984	16.884 mins
MatchAlign MGSP	26.198 secs	32	29	0.50277	1.829 hours
BEFA with Huang and Wand prior	12.667 secs	8	20	0.23109	1.609 hours
BEFA with pre-specified S	30.983 secs	8	20	0.34800	1.638 hours

As it can be seen from table 4.1, Orthogonal Procrustes is the method requiring the least time in order to solve rotational ambiguities and label switching problems. The method with the smallest normed dissimilarity between the posterior mean of the covariance matrix and the covariance matrix estimated by the posterior mean of the factor loadings matrix Λ after solving rotational ambiguities and label switching problems is Orthogonal Procrustes.

Concerning the identification of the true number of factors according to R. A. Martin et al. (2003), none of the models selecting the number of factors is able to identify it. Specifically, the Multiplicative Gamma Process Shrinkage model with MatchAlign method selects 32 factors. However, the number of active factors is equal to 6. Furthermore, BEFA with Huang and Wand Prior for the covariance matrix and BEFA with pre-specified values of S for the Inverse-Wishart prior of the covariance matrix identify 8 factors. Concerning the models with pre-specified number of factors equal or greater than the true number of factors, they are able to identify the true number of factors (R. A. Martin et al., 2003). The Classical statistics estimation of the number of factors is 4 which is the correct number of factors according to

R. A. Martin et al. (2003). Furthermore, Kaiser-Guttman criterion estimates 4 factors which is the correct number of factors according to R. A. Martin et al. (2003).

Regarding the correct allocation of the observed variables to factors according to R. A. Martin et al. (2003), for the models not selecting the number of factors, the following results indicate that the Positive Lower Triangular model with pre-specified number of factors equal to 3 merges the factor corresponding to affiliative humor and the factor corresponding to self-enhancing humor and also falsely allocates question 28 to the merged factor of affiliative and self-enhancing humor. Moreover, the Positive Lower Triangular model with pre-specified number of factors equal to 4 allocates all the questions to the correct factors except question 28 which is allocated to the factor corresponding to the affiliative humor. In addition, the Positive Lower Triangular model with pre-specified number of factors equal to 5 has the exact same results as the one with pre-specified number of factors equal to 4. The Normal Inverse-gamma model with Rotation Sign Permutation, Weighted Orthogonal Procrustes and Orthogonal Procrustes identification methods and pre-specified number of factors equal to 3 merge the factors corresponding to aggressive humor and the factor corresponding to self defeating humor. Furthermore, the model falsely allocates questions 19,23 and 31 to the factor corresponding to the affiliative humor. Additionally, the Normal Inverse-gamma model with Rotation Sign Permutation, Weighted Orthogonal Procrustes and Orthogonal Procrustes identification methods and pre-specified number of factors equal to 4 allocate correctly all the questions. The same applies to the Normal Inverse-gamma model with Rotation Sign Permutation, Weighted Orthogonal Procrustes and Orthogonal Procrustes identification methods and pre-specified number of factors equal to 5. Moreover, the Dirichlet-Laplace model with MatchAlign identification method and pre-specified number of factors equal to 3 merges the factors corresponding to affiliative and self defeating humor. In addition, the Dirichlet-Laplace model with MatchAlign identification method and pre-specified number of factors equal to 4 allocate correctly all the questions. The same applies to the Dirichlet-Laplace model with MatchAlign identification method and pre-specified number of factors equal to 5.

Regarding the correct allocation of the observed variables to factors according to R. A. Martin et al. (2003), for the models selecting the number of factors, the following results indicate that the Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the number of factors equal to $5 \log_{10} 32$ allocates all the questions correctly except ques-

tions 6 and 30 allocated together to an extra factor and question 22 allocated to another extra factor. Moreover, the BEFA model with Huang and Wand prior in S for the Inverse-Wishart prior of the covariance matrix splits almost all the correct factors. Specifically, the model splits the factor corresponding to affiliative humor into two factors allocating questions (1,13,17,21,25) to the first one and questions (5,9,29) to the second one. Additionally, the factor corresponding to self enhancing humor splits into two factors allocating the questions (2,10,14,18,26) to the first factor and questions (6,22,30) to the second one. Furthermore, the factor corresponding to aggressive humor is divided into three factors allocating questions (3,19,28) to the first one meaning that question 28 is falsely allocates to it. Similarly, questions (7,11,27) are allocated to the second factor and questions (15,23,31) to the third one respectively. Regarding the factor corresponding to self defeating humor, the model allocates all the related questions to it except the question 28. Moreover, the BEFA model with pre-specified values of S for the Inverse-Wishart prior of the covariance matrix splits the factor corresponding to affilative humor to two factors by allocating questions (5,9,17,21,29) to the first one and questions (1,13,25) to the second one. Similarly, the factor corresponding to the self-enhancing humor is divided into two factors allocating questions (2,6,14,22,30) to the first one and questions (10,18,26) to the second one. In addition, the factor corresponding to aggressive humor is split into three factors by allocating the questions (3,19,28) to the first one meaning that the model falsely allocate the 28 question to it. Also, questions (7,11,27) and questions (15,23,31) are allocated to the second and third factors respectively. Finally, the model allocates all the relevant questions to the factor corresponding to self-defeating humor, except question 28.

Diagram 4.1 shows that the Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the factors equal to $5 \log_{10} 32$ overestimates the number of factors and has the following 6 active factors: the 1st factor(1st column) corresponding to the affiliative humor, the 2nd factor(2nd column) corresponding to the self defeating humor, the 3rd factor(3rd column) corresponding to the aggressive humor, the 7th factor (7th column) corresponding to self-enhancing humor and two extra factors, them being columns 8 and 12. Furthermore, we can see that questions 6 (6th row) and 30 (30th row) are falsely allocated to the one extra factor (12th column) and question 22 (22nd row) is falsely allocated to the other extra factor (8th column).

Diagrams 4.2,4.3 and 4.4 show that the Normal Inverse-Gamma model with RSP,WOP and OP methods select the correct number of factors and allocate

the questions correctly(R. A. Martin et al., 2003) when the number of pre-specified factors is greater or equal to the true number of factors stipulated by R. A. Martin et al. (2003). Specifically, from the diagrams 4.2(c),4.3(c) and 4.4(c) it is perceived that for each model when the number of factors is equal to 5, there is a redundant factor. Consequently, as mentioned in the first chapter, we can identify that the correct number of factors is equal to 4 which is the correct number of factors (R. A. Martin et al., 2003). Additionally, it can been seen from diagrams 4.2(a),4.3(a) and 4.4(a) that the model for pre-specified number of factors equal to 3 merge the factors corresponding to aggressive humor and self defeating humor. Concerning the allocation of questions, diagrams 4.2(a),4.3(a) and 4.4(a) show that the model falsely allocates questions 19,23 and 31 to the factor corresponding to affiliative humor. Furthermore, from diagrams 4.2(b),4.3(b),4.4(b),4.2(c),4.3(c) and 4.4(c) we can see that the model allocates all the questions correctly (R. A. Martin et al., 2003). In order to make it more perceptible, an example is illustrated with the same logic also applying to the rest of the diagrams. For example, in the diagram 4.2(b) in the 1st window (1st factor) questions (3,7,11,15,19,23,27,31) are allocated correctly to the same factor(conclusively corresponding to aggressive humor), in the 2nd window (2nd factor) questions (4,8,12,16,20,24,28,32) are allocated correctly to the same factor(conclusively corresponding to self-defeating humor), in the 3rd window(3rd factor) questions (2,6,10,14,18,22,26,30) are allocated correctly to the same factor(conclusively corresponding to self-enhancing) and in the 4th window(4th factor) questions (1,5,9,13,17,21,25,29) are allocated correctly to the same factor(conclusively corresponding to affiliative humor).

Diagram 4.5 shows that the Dirichlet Laplace model with MatchAlign method allocates the questions correctly and selects the correct number of factors (R. A. Martin et al., 2003). Specifically, from diagram 4.5(c) we can perceive that there is a redundant factor. Consequently, as mentioned in the first chapter regarding the selection of the number of factors, we can identify that the correct number of factors is equal to 4 which is the correct number of factors (R. A. Martin et al., 2003). Additionally, from diagram 4.5(a), we can perceive that the first factor (first column) is the merged factor of the one corresponding to affiliative humor and the one corresponding to self-enhancing humor. Furthermore, from diagrams 4.5(b) and 4.5(c) we can see that the model allocates the observed variables correctly. In particular, in diagram 4.5(b) questions (3,7,11,15,19,23,27,31) in the 1st factor (1st column) are allocated correctly together(conclusively corresponding to aggressive humor). Similarly, questions (4,8,12,16,20,24,28,32) in the 2nd factor (2nd column) are allocated correctly together(conclusively corresponding to self-defeating

humor). Likewise, questions (2,6,10,14,18,22,26,30) in the 3rd factor (3rd column) are allocated correctly together(conclusively corresponding to self-enhancing humor). Finally, questions (1,5,9,13,17,21,25,29) in the 4th factor (4th column) are allocated correctly together(conclusively corresponding to affiliative humor).

Figure 4.1: Multiplicative Gamma Process Shrinkage model with MatchAlign identification method and initial value for the factors equal to $5 \log_{10} 32$ for Humor Style Questionnaire

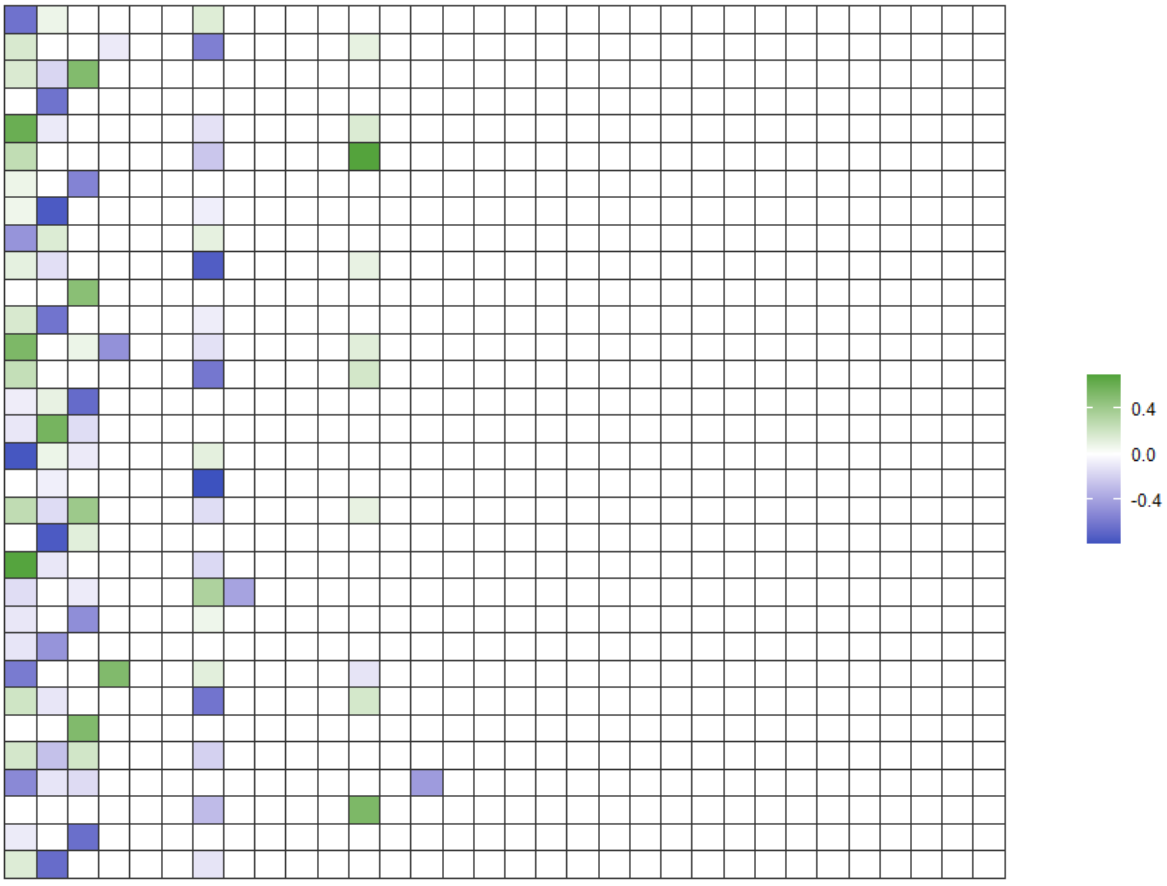
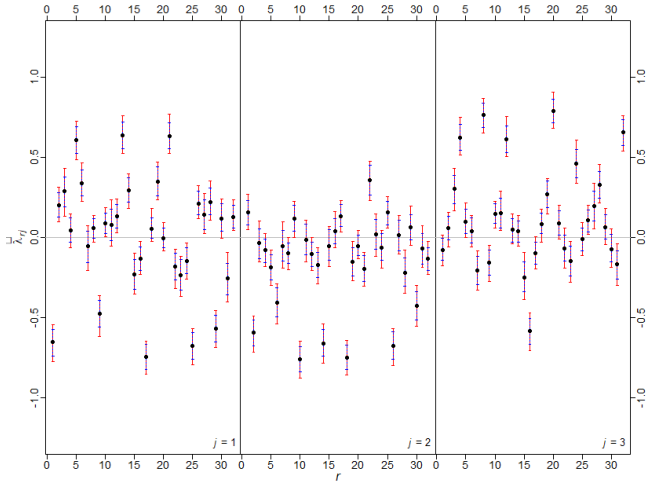
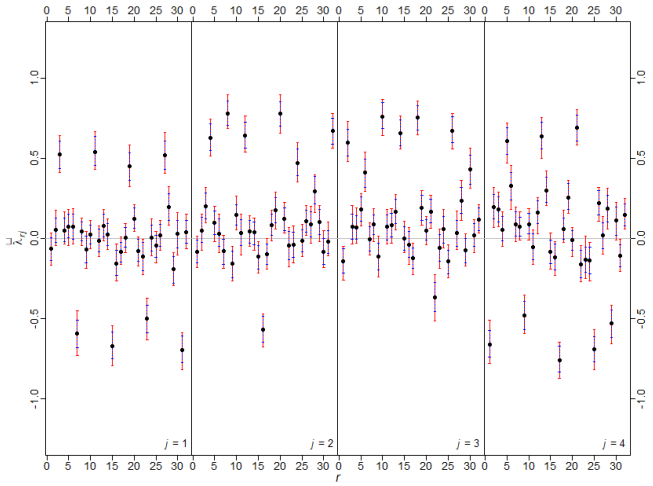


Figure 4.2: Rotation Sign Permutation With 3,4 and 5 factors for Humor Style Questionnaire

(a) Rotation Sign Permutation With 3 factors



(b) Rotation Sign Permutation With 4 factors



(c) Rotation Sign Permutation With 5 factors

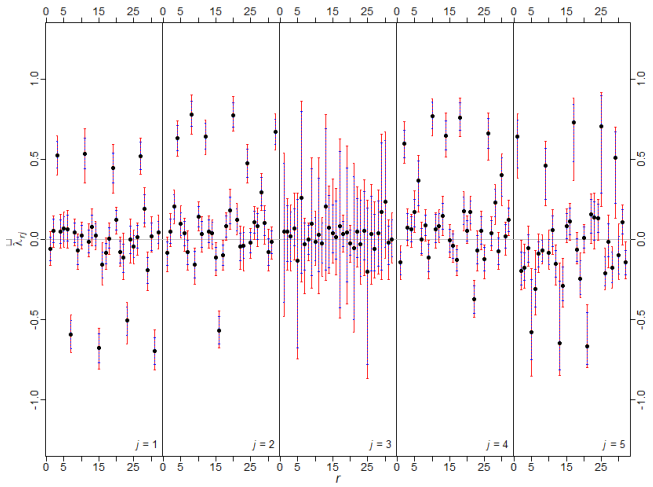
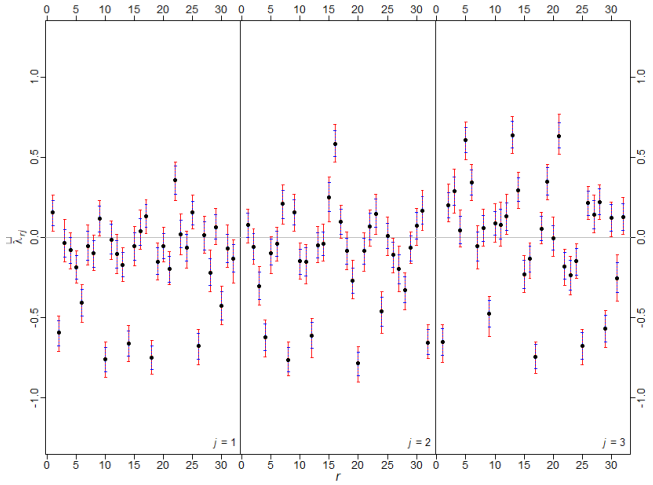
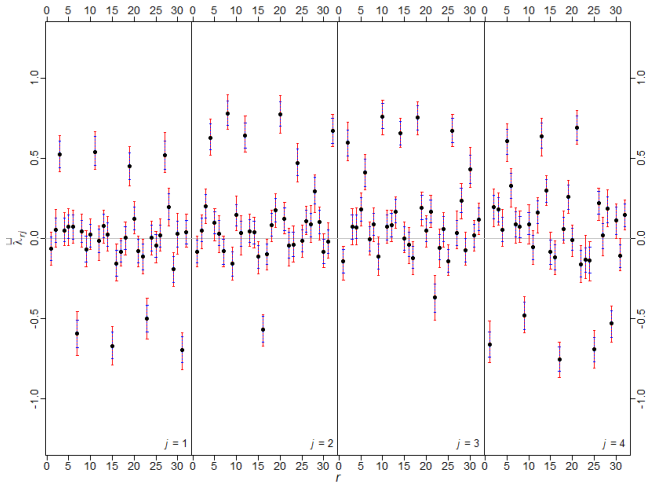


Figure 4.3: Weighted Orthogonal Procrustes With 3,4 and 5 factors for Humor Style Questionnaire

(a) Weighted Orthogonal Procrustes With 3 factors



(b) Weighted Orthogonal Procrustes With 4 factors



(c) Weighted Orthogonal Procrustes With 5 factors

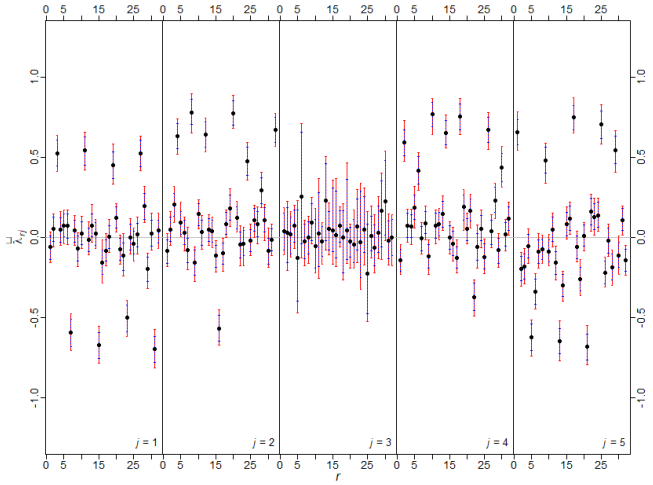
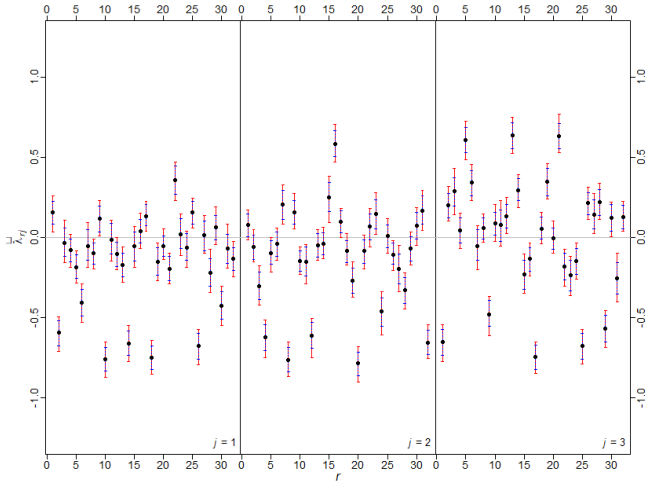
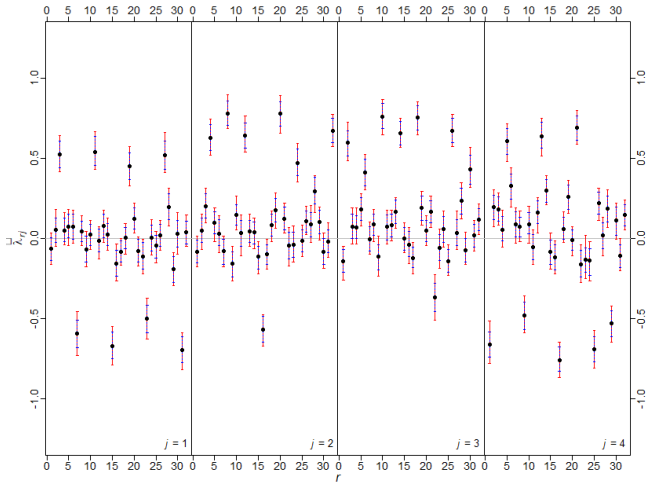


Figure 4.4: Orthogonal Procrustes With 3,4 and 5 factors for Humor Style Questionnaire

(a) Orthogonal Procrustes With 3 factors



(b) Orthogonal Procrustes With 4 factors



(c) Orthogonal Procrustes With 5 factors

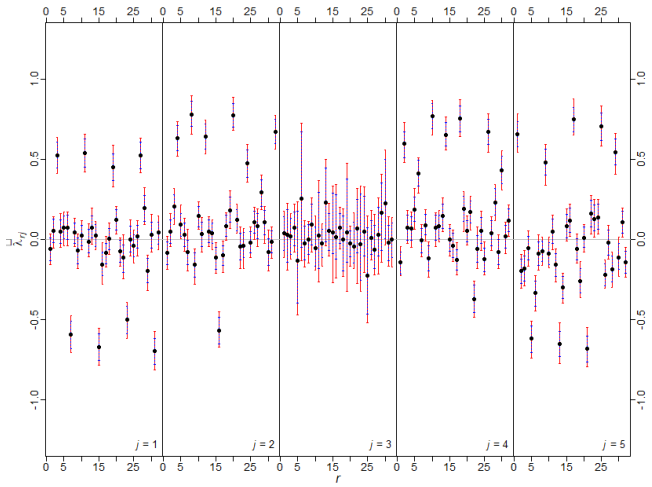
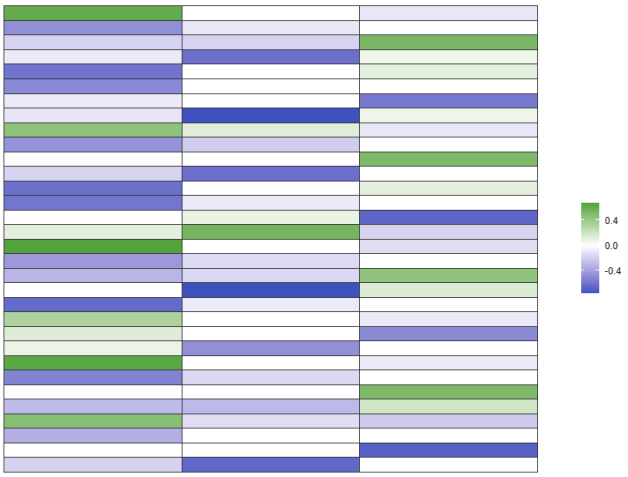
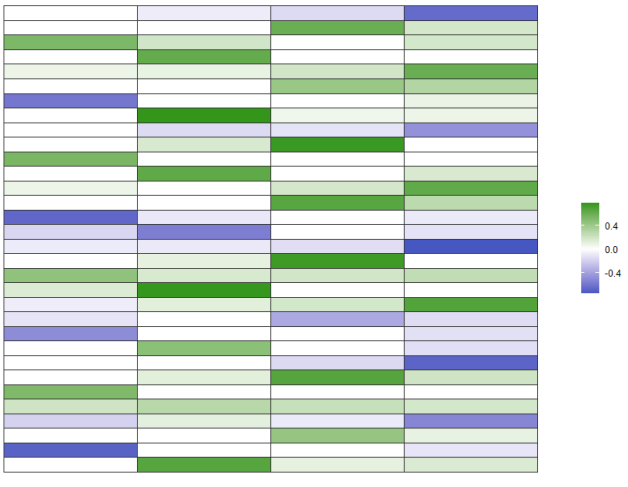


Figure 4.5: Dirichlet Laplace model with MatchAlign identification method and number of factors 3,4 and 5 for Humor Style Questionnaire

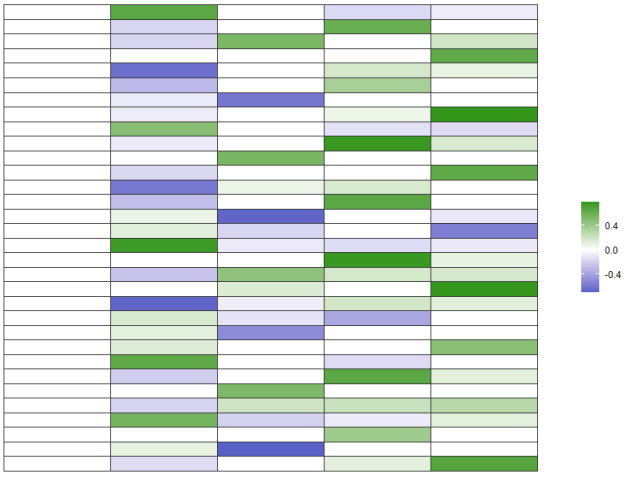
(a) MatchAlign with 3 factors



(b) MatchAlign with 4 factors



(c) MatchAlign with 5 factors



4.2 Big Five Personality Test

The Big Five Personality model was first proposed by Tupes and Christal (1958). However, it did not become well known until much later. Digman (1990) stated that personality traits can be described by the five traits of the Big Five Personality test which are Extraversion, Neuroticism, Agreeableness, Conscientiousness and Openness to Experience. According to O'Connor (2002), those five domains embody the more established personality traits and present the basic structure of a personality. Researchers investigating independently on personality traits confirmed the same five factor structure by employing different techniques and methods in order to identify the five traits, leading to different given names and interpretations for those five factors.(Norman & Goldberg, 1966; Karson & O'Dell, 1976; Goldberg, 1982; Krug & Johns, 1986; McCrae & Costa, 1987; Peabody & Goldberg, 1989; McCrae & John, 1992; Goldberg, 1992; Saucier & Goldberg, 1996; Bagby, Marshall, & Georgiades, 2005; Cattell & Mead, 2008; Costa & McCrae, 2008).

Regarding the description of each personality trait, Extraversion is associated with characteristics like friendliness, confidence, talkativeness and energeticness. People with high Extraversion love interacting with others and doing activities. Neuroticism is a personality trait associated with characteristics like unhappiness, anxiety, sullenness and emotional instability. People with high Neuroticism tend to have a pessimistic view on the world. Agreeableness associated with characteristics like trust, altruism, gentleness and love. People with high Agreeableness tend to have an optimistic view on the world. Conscientiousness is associated with self-discipline, thoughtfulness, determination and self-control. People with hight Conscientiousness prefer being organized to being spontaneous. Finally, Openness to Experience is associated with imagination, open-mindedness, creativity and self-discovery.

The data set employed can be acquired from `https://openpsychometrics.org/`. The collection of the data was made in 2012 from different online sites. The questions of the Big Five Personality Test can be found in the appendix A.2. The questionnaire has 50 questions and the responses are in a five point Likert scale varying from 1 (disagree) to 5 (agree). Each of the five personality traits is linked to 10 questions. Given that the phrases in the questions are large, the question index will be used in order to describe the links between the questions and personality traits. In particular, questions $(1, \ldots, 10)$ are linked to Extraversion personality trait, questions $(11, \ldots, 20)$ to Neuroticism , questions $(21, \ldots, 30)$ are linked to Agreeableness , questions $(31, \ldots, 40)$ are linked to Conscientiousness and questions $(41, \ldots, 50)$

are linked to Openness to Experience. Apart from the questionnaire, the respondents had to choose their race from a drop-down menu with the following available options: 1=Mixed Race, 2=Arctic (Siberian, Inuit), 3=Caucasian (European), 4=Caucasian (Indian), 5=Caucasian (Middle East), 6=Caucasian (North African, Other), 7=Indigenous Australian, 8=Native American, 9=North East Asian (Mongol, Tibetan, Korean Japanese, etc), 10=Pacific (Polynesian, Micronesian, etc), 11=South East Asian (Chinese, Thai, Malay, Filipino, etc), 12=West African, Bushmen, Ethiopian, 13=Other. Furthermore, the respondents had to state their age and choose their gender from a drop-down menu with the following options: 1(male), 2(female) and 3(other). Finally, the respondents chose whether their native language was English (1=yes, 2=no) and which of their hands they used to write (1=Right, 2=Left, 3=Both).

The full dataset consists of 19719 respondents from around the world. However, a subset will be used with respondents originating in the United States of America, consisting of 8761 respondents. Prior to the data analysis, respondents not having fully completed the questionnaire are excluded from the study. Furthermore, we check for false entries and respondents who answered non logical or non plausible answers(for example age greater than 100). After the extraction of those respondents, the subset consists of 8753 respondents with the following race characteristics 839 Mixed race, 8 Arctic, 5291 Caucasian (European), 174 Caucasian (Indian), 197 Caucasian (Middle East), 257 Caucasian (North African, Other), 3 Indigenous Australian, 139 Native American, 86 North East Asian, 20 Pacific, 341 South East Asian, 105 West African and 1204 Other. The participants' age ranges from $14-99$. There were 2945 males, 5739 females and 58 other. Regarding the allocation of native English speakers and the hand they used to write, 7899 respondents are native English speakers while 817 are not and 7652 respondents have the right hand as dominant, 815 the left hand while 247 are ambidextrous. Before proceeding with the analysis, we standardise the data and change the sign of the following questions (2,4,6,8,10,12,14,21,23,25,27,32,34,36,38,42,44,46) in order to acquire consistent answers within each personality trait.

Table 4.2: Summarized results for Big Five Personality Test

Model and identification method	Time to solve identification issues	Factors by model	Variables loaded correctly	Metric	Time to simulate MCMC samples
PLT 4 q	-	4	39	-	13.41 hours
PLT 5 q	-	5	48	-	15.198 hours
PLT 6 q	-	6	45	-	17.192 hours
RSP exact 4 q	2.402 mins	4	40	0.00439	12.498 hours
RSP exact 5 q	2.521 mins	5	50	0.00497	16.481 hours
RSP exact 6 q	6.241 mins	6	46	0.00608	18.341 hours
WOP 4 q	3.109 secs	4	40	0.00425	12.498 hours
WOP 5 q	3.410 secs	5	50	0.00456	16.481 hours
WOP 6 q	3.428 secs	6	46	0.00532	18.341 hours
OP 4 q	0.770 secs	4	40	0.00402	12.498 hours
OP 5 q	0.783 secs	5	50	0.00451	16.481 hours
OP 6 q	0.907 secs	6	46	0.00528	18.341 hours
MatchAlign DL 4 q	2.457 secs	4	40	0.00442	1.496 hours
MatchAlign DL 5 q	2.870 secs	5	50	0.00475	1.721 hours
MatchAlign DL 6 q	5.698 secs	5	50	0.00668	1.993 hours
MatchAlign MGSP	4.69 mins	49	37	0.28950	11.422 hours
BEFA with Huang and Wand prior	16.991 secs	8	40	4.99771	15.948 hours
BEFA with pre-specified S	18.392 secs	8	41	4.55831	16.094 hours

As it can be inspected from table 4.2, Orthogonal Procrustes is the method requiring the least time in order to solve rotational ambiguities and label switching problems. The method with the smallest normed dissimilarity between the posterior mean of the covariance matrix and the covariance matrix estimated by the posterior mean of the factor loadings matrix Λ after solving rotational ambiguities and label switching problems is Orthogonal Procrustes.

Concerning the identification of the true number of factors (Goldberg, 1992) for the models selecting the number of factors, the following results indicate that Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the number of the factors equal to $5 \log_{10} 50$ chooses 49 factors. However, the number of active factors is equal to 14. Additionally, both BEFA with Huang and Wand prior in S for the Inverse-Wishart prior and BEFA with pre-specified values of S for the Inverse-Wishart prior select 8 factors.

Regarding the identification of the true number of factors (Goldberg, 1992)

for the models not selecting the number of factors, the following results indicate that the Normal Inverse-Gamma with RSP, WOP and OP methods and the Positive Lower Triangular model with pre-specified number of factors equal to 4 select 4 factors. The Normal Inverse-Gamma with RSP, WOP and OP methods the Positive Lower Triangular model with pre specified number of factors equal to 5 select 5 factors which is the correct number of factors (Goldberg, 1992). The Normal Inverse-Gamma with RSP, WOP and OP methods the Positive Lower Triangular model with pre specified number of factors equal to 6 select 6 factors. The Dirichlet-Laplace model with MatchAlign method and pre-specified number of factors equal to 4 selects 4 factors. The Dirichlet-Laplace model with MatchAlign method and pre specified number of factors equal to 5 and the same with 6 identify 5 factors which is the correct number of factors (Goldberg, 1992).

The Classical statistics estimation of the number of factors is 6, which is not the correct number of factors as indicated by Goldberg (1992). However, Kaiser-Guttman criterion estimates 5 factors, which is the correct number of factors according to Goldberg (1992).

Regarding the correct allocation of the observed variables to factors (Goldberg, 1992) for the models not selecting the number of factors, the following results show that the Positive Lower Triangular model with pre-specified number of factors equal to 4 merge the factors corresponding to Neuroticism and the factor corresponding to Conscientiousness and also falsely allocates question 30 to the factor corresponding to Extraversion and questions 33 and 40 to the factor corresponding to Openness to Experience. The Positive Lower Triangular model with pre-specified number of factors equal to 5 allocates all the questions to the correct factors except question 28 which is allocated to the factor regarding Extraversion and question 34 which is allocated to Neuroticism. The Positive Lower Triangular model with pre-specified number of factors equal to 6 splits the factor corresponding to Openness to Experiences to two factors by allocating questions (41,42,44,47,48,49) to the first one, questions (43,45,46,50) to the second one and also falsely allocates question 30 to Extraversion. The Normal Inverse-Gamma model with RSP,WOP and OP methods and pre-specified number of factors equal to 4 merge the factors corresponding to Neuroticism and Conscientiousness and also falsely allocate questions 33 and 40 to the one about to Openness to Experiences and question 39 to Agreeableness. The Normal Inverse-Gamma model with RSP,WOP and OP methods and pre-specified number of factors equal to 5 allocate correctly all questions. The Normal Inverse-Gamma model with RSP,WOP and OP methods and pre-specified number of fac-

tors equal to 6 split the factor corresponding to Openness to Experience to two factors by allocating questions (43,44,45,46,49,50) to the first one and questions (41,42,47,48) to the second one. The Dirichlet-Laplace model with MatchAlign method and pre-specified number of factors equal to 4 does not allocate questions $(31, \ldots, 40)$ to any factor. The Dirichlet-Laplace model with MatchAlign method and pre-specified number of factors equal to 5 allocates correctly all questions. The same applies to the The Dirichlet-Laplace model with MatchAlign method and pre-specified number of factors equal to 6.

Regarding the correct allocation of the observed variables to factors (Goldberg, 1992), for the models selecting the number of factors, the following results stipulate that the the Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the number of factors equal to $5 \log_{10} 50$ allocates questions $(1, \ldots, 10)$ to one factor(corresponding to Extraversion), splits the factor corresponding to Neuroticism to two factors allocating questions (11,12,13,15,16,17 ,18,19) to one factor and questions (14,20) to the other one. Moreover, it splits the factor corresponding to Agreeableness to three factors allocating questions (21,22,24, 25 ,26,27,28,29) to the first one, question 23 to the second one and question 30 to the third one. It also splits the factor corresponding to Conscientiousness to two factors allocating questions (31,32,34,35,36,37,38,39,40) to the first one, and question 33 to the second one. Additionally, it splits the factor corresponding to Openness to Experience to six factors allocating questions 41 and 48 to the first one, questions 42 and 44 to the second one, questions 43 and 46 to the third one, questions 45 and 50 to the forth one, question 47 to the fifth one, and question 49 to the sixth one. Furthermore, the BEFA model with Huang and Wand prior in S for the Inverse-Wishart prior of the covariance matrix allocates correctly questions $(1, \ldots, 10)$ to one factor corresponding to Extraversion, allocates correctly questions $(11, \ldots, 20)$ to Neuroticism, allocates correctly questions $(21, \ldots, 30)$ to Agreeableness, splits the factor corresponding to Conscientiousness to two factors allocating questions (31,33,37,38,39,40) to the first factor and questions (32,34,35,36) to the second factor. Additionally, it splits the factor corresponding to Openness to Experience to three factors allocating questions (41,48,49) to the first, questions (42,44,47) to the second and questions (43,45,46,50) to the third. Moreover, the BEFA model with pre-specified values of S for the Inverse-Wishart prior of the covariance matrix allocates correctly questions $(1, \ldots, 10)$ to one factor corresponding to Extraversion, splits the factor corresponding to Neuroticism to three factors allocating questions (11,12,13) to the first factor, questions (14,17,18,20) to the second factor and questions (15,16,19) to the third factor. Furthermore,

it splits the factor corresponding to Agreeableness to two factors allocating questions (21,23,24,25,26,28,29) to the first and questions (22,27,30) to the second. It also allocates correctly questions $(31, \ldots, 40)$ to one factor corresponding to Conscientiousness and allocates correctly questions $(41, \ldots, 50)$ to Openness to Experience.

Diagram 4.6 shows that the Multiplicative Gamma Process Shrinkage model with MatchAlign method and initial value for the factors equal to $5 \log_{10} 50$ overestimates the number of factors and has the following 14 active factors: the 4th factor (4th column) corresponding to Extraversion, the 1st factor (1st column) and the 24th factor(24th column) formed after splitting the factor corresponding to Neuroticism, the 7th factor(7th column), the 8th factor(8th column) and the 34th factor (34th column) formed after splitting the factor corresponding to Agreeableness, the 5th factor(5th column) and the 12th factor (12th column) formed after splitting the factor corresponding to Conscientiousness and the 10th factor(10th column), 19th factor (19th column), 9th factor(9th column), 14th factor(14th column), 6th factor(6th column) and the 26th factor (26th column) formed after splitting the factor corresponding to Openness to Experience. Regarding the correct allocation of questions to factors (Goldberg, 1992), questions$(1, \ldots, 10)$ are allocated to factor 4 corresponding to Extraversion. The factor corresponding to Neuroticism is split to two factors allocating questions (11,12,13,15,16,17 ,18,19) to one factor(factor 1) and questions (14,20) to the other one(factor 24). Additionally, the factor corresponding to Agreeableness is split to three factors allocating questions (21,22,24,25,26,27,28,29) to the first one(factor 7), question 23 to the second one(factor 8) and question 30 to the third one(factor 34). Furthermore, the factor corresponding to Conscientiousness is split to two factors allocating questions (31,32,34,35,36,37,38,39,40) to the first one(factor 5) and question 33 to the second one(factor 20). Finally, the factor corresponding to Openness to Experience is split to six factors allocating questions 41 and 48 to the first one(factor 10), questions 42 and 44 to the second one(factor 19), questions 43 and 46 to the third one(factor 9), questions 45 and 50 to the forth one(factor 14), question 47 to the fifth one(factor 6), and question 49 to the sixth one(factor 26).

Diagram 4.7 indicates that the Dirichlet-Laplace model with MatchAlign method selects the correct number of factors and allocates the observed variables correctly (Goldberg, 1992). In particular, diagram 4.7(a) shows that some questions are not loaded onto any factor indicating that more than 4 factors are needed. Additionally, from diagram 4.7(c) we can perceive that there is a redundant factor. Consequently, as mentioned in the first chapter,

we can identify that the correct number of factors is equal to 5. Moreover, we can perceive from diagrams 4.7(b) and 4.7(c) that the model allocates the observed variables correctly (Goldberg, 1992). Specifically, questions $(1, \ldots, 10)$ are allocated to one factor corresponding to Extraversion, questions $(11, \ldots, 20)$ to Neuroticism, questions $(21, \ldots, 30)$ to Agreeableness, questions $(31, \ldots, 40)$ to Conscientiousness and questions $(41, \ldots, 50)$ to Openness to Experience

Figure 4.6: Multiplicative Gamma Process Shrinkage model with MatchAlign identification method and initial value for the factors equal to $5 \log_{10} 50$ for Big Five Personality Test

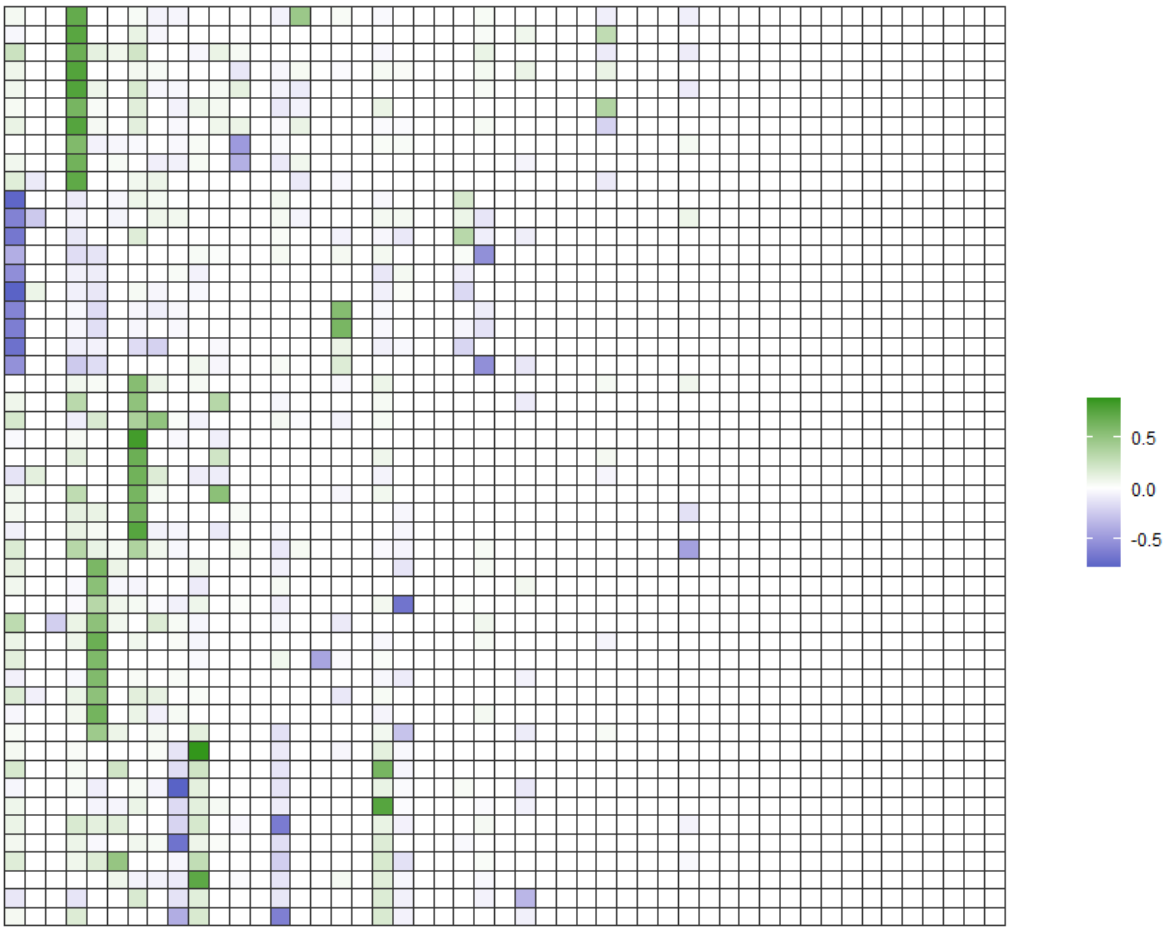
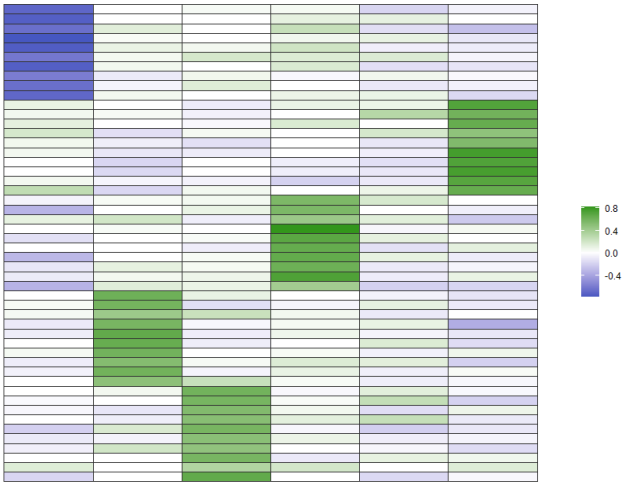


Figure 4.7: Dirichlet Laplace model with MatchAlign identification method and number of factors 4,5 and 6 for Big Five Personality Test

(a) MatchAlign with 4 factors

(b) MatchAlign with 5 factors

(c) MatchAlign with 6 factors



4.3 Final thoughts on the results of the Real data sets

Having examined the Humor Style Questionnaire and the Big Five Personality test, some remarks can be made regarding the effectiveness of each method used. As on the synthetic datasets, the Dirichlet-Laplace model with MatchAlign method is the one simulating the MCMC samples faster among the models not selecting the number of factors. Regarding the models selecting the number of factors, the findings presented in Chapter 3 were correct. Specifically, the BEFA models are faster for small numbers of observed variables, but the Multiplicative Gamma Process Shrinkage model with MatchAlign method is the better option for medium and large numbers of observed variables. Concerning the method resolving the rotational ambiguities and the label switching problem faster, the same results come out as in the synthetic data sets being achieved by Orthogonal Procrustes method. Regarding the selection of the factors, all the models selecting the number of factors are not able to identify the correct number of factors. As for the selection of the factors for the models not selecting the number of them, in the Humor Style Questionnaire data set all the models with pre-determined number of factors greater or equal to the true one are able to identify the true number of factors as stipulated by R. A. Martin et al. (2003). However, in the Big Five Personality test data set, the models were able to identify the true number of factors only when the pre-determined number of factors is equal to the true one (Goldberg, 1992). The only exception is Dirichlet-Laplace model with MatchAlign method which identifies the true number of factors for pre-specified number of factors equal or greater than the true one (Goldberg, 1992). Concerning the correct allocation of the observed variables, none of the models selecting the number of factors was able to allocate the observed variables correctly in any of the data sets. Regarding the models not selecting the number of factors, in the Humor Style Questionnaire data set all the models with pre-determined number of factors greater or equal to the true were able to allocate the observed variables correctly (R. A. Martin et al., 2003). The only exception to this is the Positive Lower Triangular model not allocating all the observed variables correctly (R. A. Martin et al., 2003). However, in the Big Five Personality test data set the models were able to allocate the observed variables correctly only when the pre-specified number of factors was equal to the true one (Goldberg, 1992). Exception to this were the Positive Lower Triangular model not allocating the observed variables correctly and the Dirichlet-Laplace model with MatchAlign method allocating the observed variables correctly according to Goldberg (1992) for

predetermined number of factor greater or equal to the true number of factors (Goldberg, 1992). All the models construct Chains reaching convergences. Regarding the metric, the method minimizing it in both data sets is Orthogonal Procrustes.

Chapter 5

Final discussions

In the first chapter, the purpose of describing the factor analysis model as well as its assumptions and its characteristics was achieved. Furthermore, the identification issue of the parameters of the Bayesian factor analysis model was explained, specifically, the uniqueness problem of the variance covariance matrix of the idiosyncratic errors and the identification issue regarding the factor loadings matrix originating by orthogonal ambiguities and label switching. Moreover, two simple Bayesian factor analysis models were presented with their assumptions and their prior specifications namely, an analogous of the Gibbs sampler employed by the MCMCpack ("MCMCpack", n.d.) package used by Papastamoulis (2018, 2020) and the model proposed by Polasek (1997). Furthermore, the issue of selecting the number of factors as well as some Classical and Bayesian methods solving it have been discussed. All the above were chosen so that the main purpose of this thesis could be comprehended in depth, as they are crucial to understanding the following chapters.

Moreover, in the second chapter, some Bayesian models were described, namely the Positive Lower Triangular method firstly proposed by Anderson and Rubin (1956), the Bayesian exploratory factor analysis method proposed by Conti et al. (2014) and the two models used with the MatchAlign method and proposed by Bhattacharya and Dunson (2011) and Bhattacharya et al. (2015). Additionally, in the same chapter, Bayesian methods solving the identification issues of the parameters of a Bayesian factor analysis model were presented including the Bayesian exploratory factor analysis method proposed by Conti et al. (2014), the Rotational Sign permutation exact, the Rotational Sign permutation Full Simulation Annealing and Rotational Sign permutation partial Simulation Annealing methods proposed by Papastamoulis and Ntzoufras (2022a), the Orthogonal procrustes and

Weight Orthogonal procrustes methods proposed by Aßmann et al. (2016) and the MatchAlign method suggested by Poworoznek et al. (2021). These models and methods were compared, contrasted and used in the practical part of this thesis, which led to some specific outcomes.

In particular, some conclusions were reached from the comparisons conducted on the three synthetic scenarios in the third chapter and the two real datasets, namely Humor Style Questionnaire and Big Five Personality test in the forth chapter.

Firstly, Dirichlet-Laplace model with MatchAlign identification method is the one simulating the MCMC samples faster among the methods not selecting the number of factors.

Secondly, Multiplicative Gamma Process Shrinkage model with MatchAlign identification method and initial value for the number of factor equal to $5log_{10}p$ requires less time to simulate MCMC samples in comparison with the BEFA model, for medium and large number of observed variables.

Additionally, the method requiring the least time in order to solve the rotational ambiguities and the label switching problem is the Orthogonal Procrustes Method.

Moreover, the Normal inverse-Gamma model with RSP,WOP,OP identification methods and the Dirichlet-Laplace model with MatchAlign identification method outperform the Positive Lower Triangular model,BEFA model and the Multiplicative Gamma Process Shrinkage model with MatchAlign identification method in the following fields based on the Humor Style Questionnaire (R. A. Martin et al., 2003) and the Big Five Personality Test (Goldberg, 1992) datasets, where the observed variables are not only correlated with the observed variables belonging to the same factor:
a) The identification of the true number of latent factors.
b) The correct allocation of the observed variables to factors.

Furthermore, the Multiplicative Gamma Process Shrinkage model overestimates the number of factors in all the experiments (see tables 3.1, 3.2, 3.3, 4.1, 4.2)

Finally, the BEFA method produces more sufficient solutions according to the metric for small and medium number of observed variables on condition that observed variables are highly correlated only with the observed variables

belonging to the same factor in comparison with the other methods (see tables 3.1 and 3.2). However, the MatchAlign method with initial value for the number of factor equal to $5log_{10}p$ produces more sufficient solutions according to the metric for large number of observed variables on condition that the observed variables are highly correlated only with the observed variables belonging to the same factor in comparison with the other methods (see table 3.3). Nevertheless, as for observed variables being correlated not only with the observed variables belonging to the same factor and for any number of observed variables, the Orthogonal Procrustes method produces more sufficient solutions according to the metric in comparison with the other methods solving rotational ambiguities and label switching (see tables 4.1 and 4.2).

One limitations of our simulated study is that in all three scenarios the variance of the idiosyncratic error is very large making the identifying of the true number of factors difficult. More scenarios could have been added were the variance smaller.

A direction for future research is to attempt to compare the models described in this thesis with the employment of different real data sets in order to evaluate the correctness of the conclusions presented. Furthermore, the comparisons can be made in different fields than the one used in order to examine different characteristics of the models and derive more conclusions regarding the efficiency of the models. Additionally, data sets with missing data or ones where the number of observed variables (p) is larger than the sample size (n) can be employed in order to examine the competence of the models under those conditions. Moreover, different hyper parameters can be used in the models described in order to evaluate whether the results coincide with the ones presented. Finally, the same comparisons as the ones made in this thesis can be conducted with models for mixtures of factor analyzers in order to evaluate the performance of more advanced Bayesian models.

Appendix A

Appendix

A.1 Bayesian Statistics

Bayesian statistics is a different way of viewing and solving statistical problems. One of the main differences with classical statistics is that in classical statistics the parameters of interest a researcher needs to estimate are unknown variables, but they have a single value. On the contrary, in Bayesian statistics the parameters of interest are random variables and the researcher must estimate which distribution those random variables follow. The same logic also applies to the parameters in Bayesian statistics where they are random variables which are characterized by a prior distribution. Another difference between the two schools of thought is that in the Bayesian statistics the researcher can use prior knowledge about the event or personal belief in the form of setting a specific prior distribution for a parameter of interest. Contrary to this, in classical statistics the modelling is based on relative frequency of an event after many times. The inference in Bayesian statistics is based on the posterior distribution of the parameters of interest. The posterior distribution is the result of the mix of the prior distribution the researcher has set and the likelihood of the observed sample.

The choice of prior in Bayesian statistics is of great importance. The researcher must specify the priors of the parameters before viewing the data and the prior distribution must have the same support as the theoretical support of the parameter. For example, if the researcher sets a prior for the variance parameter the domain of the prior distribution must be in $(0, \infty)$(for example, a researcher could use the inverse-Gamma distribution as prior distribution). Moreover, there are two different types of priors the informative priors, which have small variance most of the time and they are used when

we are very assertive about our belief and we want most of the information to be acquired from our prior and the uninformative priors or vague priors which are used when the researcher wants most of the information to be acquired from the data (Vague prior usually have larger variance). The balance between how much information the posterior will acquire from the prior and from the data is controlled from the sample size. To simplify it, when there is a small sample, most of the information is acquired from the prior and when there is a large sample, most of the information is acquired from the data. For the majority of the occasions it is optimal to use vague priors.

A.1.1 Bayes Theorem

Lets assume $(B_1, B_2, \ldots, B_n)$ a partition of the sample space Ω such as $P(B_i) > 0$ for every $i = 1, 2, \ldots, n$ then for every probability A of the same sample space with $P(A) > 0$ it is true that

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)}$$
$$= \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$
$$\text{for } i = 1, 2, \ldots, n.$$

Partition of the sample space is when we have $\Omega = B_1 \cup B_2 \cup \cdots \cup B_n$ and $B_i \cap B_j = \emptyset$ for every $i \neq j$.

The above equation can be rewritten in the following form. Lets denote our observed data as y and the likelihood of our data as $f(y|\theta)$ where θ is the vector of the parameters. Lets also denote as $\pi(\theta)$ the prior distributions we have set for our parameters and as $f(y)$ the marginal distribution of the data and also as $f(\theta|y)$ the posterior distribution of our parameter. Now we can rewrite the Bayes rule in the following way

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}$$

where the likelihood $f(y|\theta)$ is the information we acquire from our observed sample and has the following form

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

Now for the denominator of the Bayes rule, if $\pi(\theta)$ is absolute continuous, it can be written as

$$f(y) = \int_{\Theta} f(y|\theta)\pi(\theta)d\theta$$

and if $\pi(\theta)$ is discrete, it can be written as

$$f(y) = \sum_{\Theta} f(y|\theta)\pi(\theta).$$

A.1.2 Gibbs Sampler

Gibbs sampler is one of the most well known MCMC techniques and it is a special case of the Metropolis Hastings algorithm which will be discussed in the next subsection. It was created by Geman and Geman (1984). In order for the algorithm to be implemented properly, the researcher must have knowledge about the conditional nature of the relationship between the variables of interest. The main principle of the algorithm is that if it is possible to specify each of the variables as conditioned on the others, then by iterating those conditional distributions, the algorithm converges to the true joint distribution. The Gibbs sampler algorithm works as specified in the following paragraph.

Lets assume that the distribution of interest is $f(\theta)$ where θ is the vector of parameters with length k and our objective is to estimate the posterior distribution of those parameters θ. Lets also denote as Θ the full conditional distribution of θ and define it as $f(\Theta) = f(\theta_i|\theta_{-i})$ for $i = 1, 2, \ldots, k$. The notation θ_{-i} means all the other parameters except the parameter θ_i. Also, T is the number of iterations for the Gibbs sampler algorithm. An important note we must make about the Gibbs sampler is that the fully conditional distribution of each parameter in the θ vector must be analytically defined and also that there is a way of sampling from this conditional distribution

in order to implement the Gibbs sampler. In order to implement a Gibbs sampler, one must follow the following steps

1. Choose starting values $\Theta^{[0]} = (\theta_1^{[0]}, \theta_2^{[0]}, \ldots, \theta_k^{[0]})$
2. Update for $i = 1, 2, \ldots, T$
   $\theta_1^{[i]} \sim f(\theta_1 | \theta_2^{[i-1]}, \theta_3^{[i-1]}, \ldots, \theta_{k-1}^{[i-1]}, \theta_k^{[i-1]})$
   $\theta_2^{[i]} \sim f(\theta_2 | \theta_1^{[i]}, \theta_3^{[i-1]}, \ldots, \theta_{k-1}^{[i-1]}, \theta_k^{[i-1]})$
   $\vdots$
   $\theta_{k-1}^{[i]} \sim f(\theta_{k-1} | \theta_1^{[i]}, \theta_2^{[i]}, \ldots, \theta_{k-2}^{[i]}, \theta_k^{[i-1]})$
   $\theta_k^{[i]} \sim f(\theta_k | \theta_1^{[i]}, \theta_2^{[i]}, \ldots, \theta_{k-2}^{[i]}, \theta_{k-1}^{[i]})$

If the Gibbs sampler has run for a sufficient number of times and has achieved convergence, then all future iterations from this point produce samples from the posterior distribution.

A.1.3 Metropolis Hastings

Metropolis Hastings is one of the most well known MCMC techniques. The genesis of the algorithm can be found in Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953). The description of the Metropolis Hastings algorithm with a single parameter vector is the following: assuming that $\theta \in \Theta^k$ is a k length vector of parameters we have to estimate and $\pi(\theta)$ is the posterior distribution of interest, lets denote as $q_t(\theta'|\theta)$ the proposal distribution which in each step of the Markov chain suggests a new move from θ to θ'. The choice of q_t must be made so that it can fit some criteria. The first criterion is that it must have the same support as the theoretical support of the parameter. The second criterion is that it is simple to sample from this distribution. The last criterion is the fraction $\frac{\pi(\theta')}{q_t(\theta'|\theta)}$ fully known up to some arbitrary constant, independent of θ. Another important aspect for the choice of the proposal distribution is the variance of the distribution. If a researcher chooses large variance, then the sampler will move too far in each successive step causing the sampler to explore the sample space in exaggerated steps, which leads to slow convergences and poor mixing. If a researcher chooses small variance, then the sampler will move too close in each successive step causing the sampler to require a large amount of steps

in order to explore the sample space which also leads to slow convergences and poor mixing. Because of this way of generating samples (the new move is conditional on the last move) there are dependencies in both the posterior and the candidate generating form. An important, note we need to make before proceeding with the description is that in order for the algorithm to work properly, it must be possible to calculate the reverse function $q_t(\theta|\theta')$. Now lets, define as acceptance ratio the following fraction

$$a(\theta', \theta) = \frac{\pi(\theta')q_t(\theta|\theta')}{\pi(\theta)q_t(\theta'|\theta)}.$$

In each iteration of the algorithm the decision to move from a value θ to a new value θ' is probabilistically determined according to

$$\theta^{[t+1]} = \begin{cases} \text{move to} \quad \theta' & \text{with probability } \min(a(\theta', \theta^{[t]}), 1) \\ \text{stay at} \quad \theta & \text{with probability 1 - } \min(a(\theta', \theta^{[t]}), 1) \end{cases}.$$

Having described the essential aspects about the algorithm, one step from the Metropolis Hastings algorithm will be described as follow

1. sample $\pi(\theta')$ from the proposal distribution $q_t(\theta'|\theta)$ where θ is the current location and θ' is the proposed new location.
2. sample u from a uniform distribution $u \sim U[0, 1]$.
3. If $a(\theta', \theta) = \frac{\pi(\theta')q_t(\theta|\theta')}{\pi(\theta)q_t(\theta'|\theta)} > u$
   then update $\theta = \theta'$
   else keep θ.

Each step of the algorithm can have two different outcomes in three different ways. The first outcome of the algorithm can be updating θ to θ' and this outcome can be achieved in two different ways. The first way is by sampling a value of higher density and updating θ to θ' and the second way is by sampling a value of relative small density, but also sampling a small variable from the uniform distribution. The second outcome is staying in the old θ which can by achieved by sampling a value of relative small density and also sampling a large variable from the uniform distribution. Therefore,

the Metropolis Hastings algorithm does not always move to higher density points. On the contrary it describes the full posterior density by moving to high density points and also low density points allowing the algorithm to discover the full posterior density.

The convergence of the Metropolis Hastings algorithm to the limiting distribution of $\pi(\theta)$ can be attributed to the detailed balance equation and some very general conditions. Before we explain the detailed balance equation, we need to define some components of this equation. Lets define the transition kernel from the chain point θ to the chain point θ' as $p(\theta, \theta')$. Taking into account that Metropolis Hastings algorithm has the reversibility condition, lets define as $p(\theta', \theta)$ the transition kernel from the chain point θ' to the chain point θ. An important note about the Metropolis Hastings algorithm transition kernel is that it can be broken down to two parts : the jumping density $q(\theta'|\theta)$ and a jumping probability $d(\theta, \theta')$ and has the attribute of being bound from the limiting distribution of $\pi(\theta)$. Lets also define as $\pi(\theta)$ and $\pi(\theta')$ the value of the posterior distribution for the chain point θ and θ'. The detailed balance equation is the following

$$\pi(\theta)p(\theta, \theta') = \pi(\theta')p(\theta', \theta).$$

Robert and Casella (2004) proved that under very general conditions and provided that the balance equation holds then any distribution with the proper support is good proposal distribution that contributes a Metropolis Hastings chain, which will converge to the limiting distribution of $\pi(\theta)$ if we let the algorithm be executed for a sufficient number of iterations.

An important difference between the Gibbs sampler and Metropolis Hastings is that Gibbs sampler makes a move in each iteration in contrast to Metropolis Hastings which does not make a move in each iteration.

A.1.4 Empirical convergences diagnostics

Geweke Time Series Diagnostic created by J. Geweke (1992) is a difference of means test comparing the difference between the mean calculated in a proportion of the begging of the chain after the burn in and the mean in a proportion in the last part of the chain. An important note about the proportions is that they must not be overlapping. The key idea of this method

is that as the difference of means test follows asymptotically the standard normal distribution, when the difference between the two means is a value which is irregular for a standard normal distribution. This will indicate that the two proportions differ greatly. Consequently the chain has not converged. The null hypothesis of the test is that the chain has been converged. In order to define the test statistic, we need first to define its components. Lets define as θ_1 with length n_1 the proportion from the beginning of the chain after the burn in and as θ_2 with length n_2 the proportion from the last part of the chain. J. Geweke (1992) suggests the size of the proportion should satisfy the ratios $\frac{n_1}{n} = 0.1$ for the first proportion and $\frac{n_2}{n} = 0.5$ for the second, where n is the size of the chain after the burn in. The choice of those ratios is made in order to encapsulate the 0 to 0.1 quantiles of the chain and the 0.5 to 1 quantiles of the chain. Lets also define as g the function of interest which for the majority of times is the mean in each proportion and define as $s_1(0)$ and $s_2(0)$ the symmetric spectral density function with the assumption that there are no discontinuities at 0. Having defined the components of the Geweke Time Series Diagnostic test, its statistical equation is the following

$$G = \frac{\bar{g}(\theta_1) - \bar{g}(\theta_2)}{\sqrt{\frac{s_1(0)}{n_1} + \frac{s_2(0)}{n_2}}}.$$

Geweke Diagnostic G converges in standard normal under the null hypothesis and by keeping the proportions fixed and letting the MCMC algorithm run for a sufficient amount of iterations. The last part is needed in order for the central limiting theorem to be applied. A rule of thump for the Geweke Diagnostic is that any value of G greater than 2 is a sign of non convergence.

Gelman and Rubins Multiple Sequence Diagnostic created by Gelman and Rubin (1992) is a test which compares chains with non-identical starting positions which are overdispersed. Most of the times 5 to 10 chains are sufficient, but more complicated models may need more chains. For the choice of the overdispersed starting point, Gelman and Rubin (1992) proposed the use of EM algorithm from different starting points in order to find modes and use them as starting points for the MCMC algorithm. The inspiration for this method derives from the ANOVA based tests and normal theory approximations to the marginal posteriors. The philosophy of this method stipulates that before the chains have reached convergence to the target distribution, the within chain variance underestimates the total posterior variance in θ. Contrary to this the estimated variance overestimates the total posterior vari-

ance in θ because we have chosen starting positions which are overdispersed in comparison with the target distribution. However, when the chains have been converged, the difference between those two estimations is negligible. An important note about the chains is that they must have the same size. The steps of the algorithm for each parameter of interest are interpreted as follows.

1. Execute the MCMC algorithm for $m \geq 2$ chains with each chain having size equal to $2n$ and be initialized from overdispersed starting positions. The chains will have the following form

$$\begin{array}{cccccccc}
\text{chain 1} & \theta_{(1)}^{[0]} & \theta_{(1)}^{[1]} & \dots & \theta_{(1)}^{[n]} & \dots & \theta_{(1)}^{[2n-1]} & \theta_{(1)}^{[2n]} \\
\text{chain 2} & \theta_{(2)}^{[0]} & \theta_{(2)}^{[1]} & \dots & \theta_{(2)}^{[n]} & \dots & \theta_{(2)}^{[2n-1]} & \theta_{(2)}^{[2n]} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{chain m-1} & \theta_{(m-1)}^{[0]} & \theta_{(m-1)}^{[1]} & \dots & \theta_{(m-1)}^{[n]} & \dots & \theta_{(m-1)}^{[2n-1]} & \theta_{(m-1)}^{[2n]} \\
\text{chain m} & \theta_{(m)}^{[0]} & \theta_{(m)}^{[1]} & \dots & \theta_{(m)}^{[n]} & \dots & \theta_{(m)}^{[2n-1]} & \theta_{(m)}^{[2n]}
\end{array}$$

   discard the first n iterations as burn in period.

2. Calculate the following values for each parameter

   I) Within Chain Variance

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{(j)}^{[i]} - \bar{\theta_j})^2$$

   where $\bar{\theta_j}$ denotes the mean of the chain j after the burn in.

   II) Between Chain Variance

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta_{(j)}} - \bar{\bar{\theta}})^2$$

   where $\bar{\bar{\theta}}$ denotes the mean of means.

   III) Estimated Variance

$$\widehat{Var}(\theta) = (1 - \frac{1}{n})W + (\frac{1}{n})B.$$

3. Calculate the Shrink Factor which is the convergence diagnostic and has one value for each parameter of interest

$$\hat{R} = \sqrt{\frac{\widehat{Var}(\theta)}{W}}.$$

4. Examine if the values of $\hat{R}$ are less than 1.2

$\hat{R} \leq 1.2$ than the m chains have all converged to the target distribution.

$\hat{R} > 1.2$ some of the chains have not converged to the target distribution.

Heidelberger and Welch Diagnostic was created by Heidelberger and Welch (1981). Its primary objective was the simulation in operation research. Its conception came from Browhian bridge theory (Revuz & Yor, 2013) and Cramer-von Mises statistic(Cramér, 1928). The Null hypothesis of this test is that the chain has reached convergence to the stationary distribution. An important note about this method is that each parameter is tested separately. Before explaining the algorithm and the test statistic of the method, its component should be defined. Defining as T the length of the chain after the burn in and the discard(We will explain more about the discard in the algorithm) and the proportion of the chain at test as s, which takes values between 0 and 1. Lets denote as $\lfloor x \rfloor$ the floor function(for example $\lfloor 4.7 \rfloor = 4$) and define as $T_{\lfloor sT \rfloor} = \sum_{i=1}^{\lfloor sT \rfloor} \theta_i$ the sum of the values of the chain from the beginning of the chain up to the value below sT. Lets define as $s(0)$ the spectral density of the chain and as $\bar{\theta}$ the mean of the chain. The Heidelberger and Welch Diagnostic test statistic has the following form

$$B_T(s) = \frac{T_{\lfloor sT \rfloor} - \lfloor sT \rfloor \bar{\theta}}{\sqrt{Ts(0)}}.$$

The algorithm has two parts.

PART 1

A1 Specify the number of iterations (N), the accuracy (ε) and the statistical significance(α).

A2 Calculate the test statistic($B_T(s)$) using 10% of the chain iterations.

A3 If the test rejects the null hypothesis, discard the first 10% of the chain iterations and execute the test again for the next 10% .

A4 This procedure will terminate if 50% of the chain has been discarded or the test fails to reject the null hypothesis in an iteration.

PART 2 If a part of the data passes the test, then the halfwidth analysis part is executed to the part of the data not discarded.

B1 Calculate the halfwidth $(1-\alpha)\%$ credible interval around the sample mean. The estimated asymptotic standard error is calculated as $\sqrt{\frac{s(0)}{n^*}}$ where n^* is the part of the data which is not discarded.

B2 Examine if the mean divided by the halfwidth is less than ε

If True The test is passed for this parameter.

If False The researcher needs to execute more MCMC iterations in order to empirically estimate the mean with accuracy as well as for the chain to reach convergence.

The Raftery and Lewis Integrated Diagnostic was created by Raftery and Lewis (1996). This method's primary objective is to estimate the appropriate chain length in order to have efficient results. Therefore, this method estimates the sufficient length of the burn-in period, the thinning and the number of iterations in order to estimate the posterior quantities of interest appropriately. Before explaining how this method operates, some of its components should be defined. Let's define as q the posterior quantile of interest, as r the acceptable tolerance for the quantile q and as s the desired probability of being within that tolerance, as $\Phi()$ the cumulative distribution function of the Standard Normal distribution and as e the convergence tolerance which determines the stopping point according to the parallel chain. A typical choice for the convergence tolerance is the value 0.001. This method requires the execution of two chains for each parameter of interest, one primary chain and one parallel chain which is not a Markov chain and its length is equal to n_{pilot}. The values of the parallel chain are binary and the value in each iteration is determined according to whether the generated values in the primary chain up to this point are less or greater than the chosen quantile. When this procedure ends, the parallel chain acquires some Markovian properties allowing it to estimate the primary chain probabilities that satisfy the researcher constraints. An important note about this method is that each implementation estimates the appropriate length of the chain for only one parameter of interest. The algorithm of the Raftery and Lewis Integrated Diagnostic is the following

1. Select the posterior quantile of interest q, the acceptable tolerance for this quantile r, the desired probability of being within the tolerance s and the convergence tolerance e.

2. Run the two chains with the parallel chain having length equal to

$$n_{pilot} = \left[\Phi^{-1}(\frac{s+1}{2}) \frac{\sqrt{q(1-q)}}{r} \right]^2 .$$

3. Run the new chain with length of the burn in equal to the estimated. The length of the thinning equal to the estimated. Finally, the number of iterations after the burn in equal to the estimated.

4. Use the new chain to estimate the parameter of interest.

A.2 Real Data sets questionnaires

Table A.1: Humor Styles Questionnaire

Answer / Question	Never or very rarely true	Rarely true	Sometimes true	Often true	Very often or always true
I usually do not laugh or joke around much with other people					
If I am feeling depressed, I can usually cheer myself up with humor					
If someone makes a mistake, I will often tease them about it					
I let people laugh at me or make fun at my expense more than I should					
I do not have to work very hard at making other people laugh I seem to be a naturally humorous person					
Even when I am by myself, I am often amused by the absurdities of life.					
People are never offended or hurt by my sense of humor					
I will often get carried away in putting myself down if it makes my family or friends laugh					

I rarely make other people laugh by telling funny stories about myself					
If I am feeling upset or unhappy I usually try to think of something funny about the situation to make myself feel better					
When telling jokes or saying funny things, I am usually not very concerned about how other people are taking it					
I often try to make people like or accept me more by saying something funny about my own weaknesses, blunders, or faults					
I laugh and joke a lot with my closest friends					
My humorous outlook on life keeps me from getting overly upset or depressed about things					
I do not like it when people use humor as a way of criticizing or putting someone down					
I do not often say funny things to put myself down					
I usually do not like to tell jokes or amuse people					
If I am by myself and I am feeling unhappy, I make an effort to think of something funny to cheer myself up					
Sometimes I think of something that is so funny that I can not stop myself from saying it, even if it is not appropriate for the situation					
I often go overboard in putting myself down when I am making jokes or trying to be funny					
I enjoy making people laugh					
If I am feeling sad or upset, I usually lose my sense of humor					
I never participate in laughing at others even if all my friends are doing it					
When I am with friends or family, I often seem to be the one that other people make fun of or joke about					
I do not often joke around with my friends					
It is my experience that thinking about some amusing aspect of a situation is often a very effective way of coping with problems					
If I do not like someone, I often use humor or teasing to put them down					
If I am having problems or feeling unhappy, I often cover it up by joking around, so that even my closest friends do not know how I really feel					
I usually can not think of witty things to say when I am with other people					

I do not need to be with other people to feel amused – I can usually find things to laugh about even when I am by myself					
Even if something is really funny to me, I will not laugh or joke about it if someone will be offended					
Letting others laugh at me is my way of keeping my friends and family in good spirits					

Table A.2: Big 5 Personality Test Questionnaire

Answer / Question	Disagree	Moderately Disagree	Neutral	Moderately Agree	Agree
I am the life of the party					
I do not talk a lot					
I feel comfortable around people					
I keep in the background					
I start conversations					
I have little to say.					
I talk to a lot of different people at parties					
I do not like to draw attention to myself					
I do not mind being the center of attention					
I am quiet around strangers					
I get stressed out easily					
I am relaxed most of the time					
I worry about things					
I seldom feel blue					
I am easily disturbed					
I get upset easily					
I change my mood a lot					
I have frequent mood swings					
I get irritated easily					
I often feel blue					
I feel little concern for others					
I am interested in people					
I insult people					
I sympathize with others feelings					
I am not interested in other peoples problems					
I have a soft heart					
I am not really interested in others					
I take time out for others					
I feel others emotions					
I make people feel at ease					
I am always prepared					
I leave my belongings around					
I pay attention to details					

I make a mess of things					
I get chores done right away					
I often forget to put things back in their proper place					
I like order					
I shirk my duties					
I follow a schedule					
I am exacting in my work					
I have a rich vocabulary					
I have difficulty understanding abstract ideas					
I have a vivid imagination					
I am not interested in abstract ideas					
I have excellent ideas					
I do not have a good imagination					
I am quick to understand things					
I use difficult words					
I spend time reflecting on things					
I am full of ideas					

A.3 Distributions

Multivariate Normal distribution

The Multivariate Normal distribution is a continuous multivariate distribution and its probability density function is the following

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

The Multivariate Normal distribution with location parameters μ and covariance Σ has mean equal to the vector μ, covariance equal to the matrix Σ which is a positive semi defined matrix and support on R^p

Uniform distribution

The Uniform distribution is a continuous distribution and its probability density function is the following

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a \text{ and } x > b \end{cases}$$

The uniform distribution has mean equal to $\frac{1}{2}(a+b)$, variance equal to $\frac{1}{12}(b-a)^2$ and support on $[a, b]$

Gamma distribution

The Gamma distribution is a continuous distribution and its probability density function is the following

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

The Gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$ has mean equal to $\frac{a}{b}$, variance equal to $\frac{a}{b^2}$ and support on $(0, +\infty)$

Inverse Gamma distribution

The Inverse Gamma distribution is a continuous distribution and its probability density function is the following

$$f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-\frac{b}{x}}$$

The Inverse Gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$ has mean equal to $\frac{b}{a-1}$ with the restriction that $a > 1$, variance equal to $\frac{b^2}{(a-1)^2(a-2)}$ with the restriction that $a > 2$ and support on $(0, +\infty)$

Exponential distribution

The Exponential distribution is a continuous distribution and its probability density function is the following

$$f(x) = \lambda e^{-\lambda x}$$

The Exponential distribution with rate parameter $\lambda > 0$ has mean equal to $\frac{1}{\lambda}$, variance equal to $\frac{1}{\lambda^2}$ and support on $[0, +\infty)$

Double Exponential or Laplace distribution

The Double Exponential or Laplace distribution is a continuous distribution and its probability density function is the following

$$f(x) = \frac{1}{2b} \exp(-\frac{|x - \mu|}{b})$$

The Double Exponential or Laplace distribution with location parameter μ and scale parameter $b > 0$ has mean equal to μ, variance equal to $2b^2$ and

support on R

Dirichlet distribution

Dirichlet distribution is a continuous multivariate distribution and its probability density function is the following

$$f(x) = \frac{1}{B(a)} \prod_{i=1}^{K} x_i^{a_i - 1}$$

$$\text{where } B(a) = \frac{\prod_{i=1}^{K} \Gamma(a_i)}{\sqrt{a_0}}$$

$$\text{where } a_0 = \sum_{i=1}^{K} a_i$$

The Dirichlet distribution with integer number of categories $K \geq 2$ and concentration parameters $a = (a_1, \ldots, a_k)$ with $a_i > 0$ has mean equal to $\frac{a_i}{a_0}$, variance equal to $\frac{\tilde{a}_i(1-\tilde{a}_i)}{a_0+1}$, covariance equal to $\frac{\delta_{ij}\tilde{a}_i - \tilde{a}_i\tilde{a}_j}{a_0+1}$ where δ_{ij} denotes the Kronecker delta and $\tilde{a}_i = \frac{a_i}{a_0}$ and support for each x_i in $[0, 1]$ but with $\sum_{i=1}^{K} x_i = 1$

Generalized Inverse Gaussian distribution

Generalized Inverse Gaussian distribution is a continuous distribution and its probability density function is the following

$$f(x) = \frac{(\frac{a}{b})^{\frac{p}{2}}}{2K_p(\sqrt{ab})} x^{(p-1)} \exp\left(\frac{-(ax + \frac{b}{x})}{2}\right)$$

The Generalized Inverse Gaussian distribution with parameters $(a > 0, b > 0, p \in R)$ has mean equal to $\frac{\sqrt{b}K_{p+1}(\sqrt{ab})}{\sqrt{a}K_p(\sqrt{ab})}$, variance equal to $(\frac{b}{a})\left(\frac{K_{p+2}(\sqrt{ab})}{K_p(\sqrt{ab})} - \left(\frac{K_{p+1}(\sqrt{ab})}{K_p(\sqrt{ab})}\right)^2\right)$ and support on $(0, +\infty)$

Beta distribution

Beta distribution is a continuous distribution and its probability density function is the following

$$f(x) = \frac{x^{(a-1)}(1-x)^{(b-1)}}{B(a, b)}$$

$$\text{Where } B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

The beta distribution with shape parameter $a > 0$ and scale parameter $b > 0$ has mean equal to $\frac{a}{a+b}$, variance equal to $\frac{ab}{(a+b)^2(a+b+1)}$ and support on $[0,1]$

Inverse Wishart distribution

Inverse Wishart distribution probability density function is the following

$$f(x) = \frac{|\Psi|^{\frac{v}{2}}}{2^{v\frac{p}{2}}\Gamma_p(\frac{v}{2})}|x|^{\frac{-(v+p+1)}{2}}\exp(-\frac{1}{2}Tr(\Psi x^{-1}))$$

Where (x,Ψ) denotes positive definite matrices of $p \times p$ dimensions, $||$ denotes the determinant and $\Gamma_p()$ denotes the multivariate gamma function.

The inverse Wishart distribution with degrees of freedom equal to $v > p-1$ and scale matrix Ψ which is positive defined has mean equal to $\frac{\Psi}{v-p-1}$, variance equal to $\frac{(v-p+1)\Psi_{ij}^2+(v-p-1)\Psi_{ii}\Psi_{jj}}{(v-p)(v-p-1)^2(v-p-3)}$ and covariance equal to $\frac{2\Psi_{ij}\Psi_{kl}+(v-p-1)(\Psi_{ik}\Psi_{jl}+\Psi_{il}\Psi_{kj})}{(v-p)(v-p-1)^2(v-p-3)}$

Wishart distribution

Wishart distribution probability density function is the following

$$f(x) = \frac{|x|^{\frac{(n-p-1)}{2}}\exp\left(\frac{-trace(V^{-1}x)}{2}\right)}{2^{\frac{np}{2}}|V|^{\frac{n}{2}}\Gamma_p(\frac{n}{2})}$$

Where (x,V) denotes positive definite matrices of $p \times p$ dimensions and $\Gamma_p()$ denotes the multivariate gamma function.

The Wishart distribution with degrees of freedom equal to $v > p-1$ and scale matrix $V > 0$ which is positive defined has mean equal to nV, variance equal to $n(u_{ij}^2 + u_{ii} + u_{jj})$

Truncated Normal distribution

Truncated Normal distribution is a continuous distribution and its probability density function is the following

$$f(x) = \frac{1}{\sigma}\frac{\varphi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

Where $\varphi(x)$ denotes the probability density function of the standard normal distribution and $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution

The Truncate Normal distribution with minimum value of x equal to a, maximum value of x equal to b, $\mu \in R$ and $\sigma^2 > 0$ has mean equal to $\mu + \frac{\varphi(\frac{a-\mu}{\sigma})-\varphi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}\sigma$, variance equal to $\sigma^2\left(1-\frac{\frac{b-\mu}{\sigma}\varphi(\frac{b-\mu}{\sigma})-\frac{a-\mu}{\sigma}\varphi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}-\left(\frac{\varphi(\frac{a-\mu}{\sigma})-\varphi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}\right)^2\right)$ and support on $[a, b]$

Poisson distribution

Poisson distribution is a discrete distribution and its probability mass function is the following

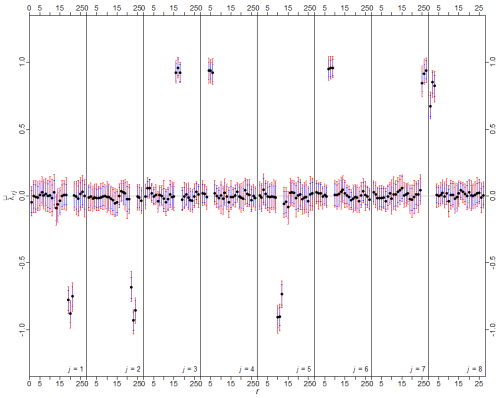
$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

The Poisson distribution with rate parameter $\lambda > 0$ has mean equal to λ variance equal to λ and support on the Natural numbers $x \in N$

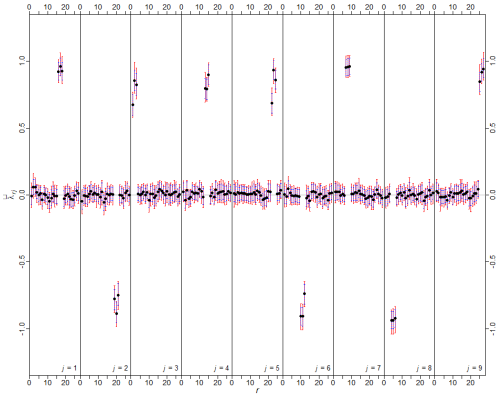
A.4 Plots

Figure A.1: Weight Orthogonal Procrustes With 8,9,10 factors for scenario 2

(a) Weight Orthogonal Procrustes with 8 factors



(b) Weight Orthogonal Procrustes with 9 factors



(c) Weight Orthogonal Procrustes with 10 factors

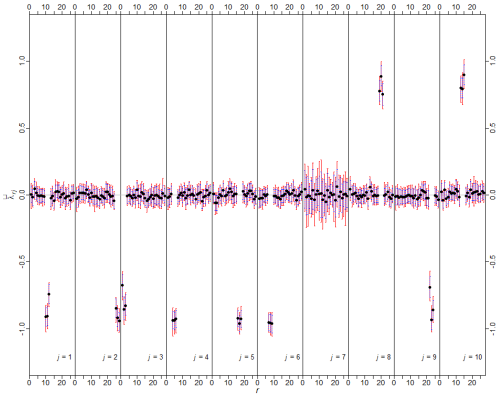
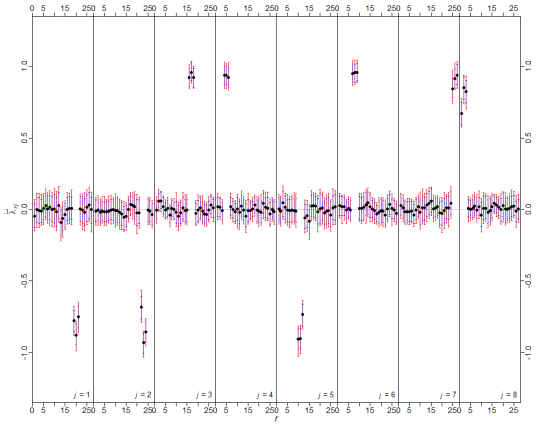
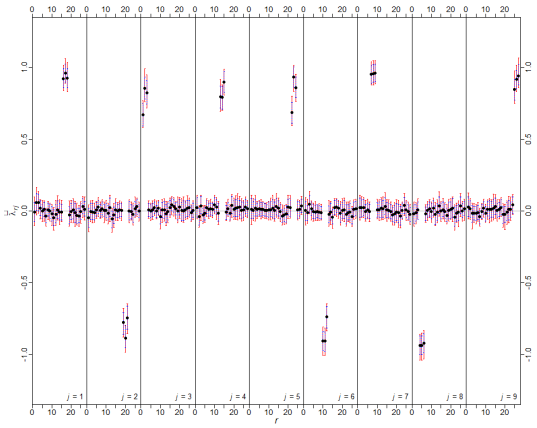


Figure A.2: Orthogonal Procrustes With 8,9,10 factors for scenario 2

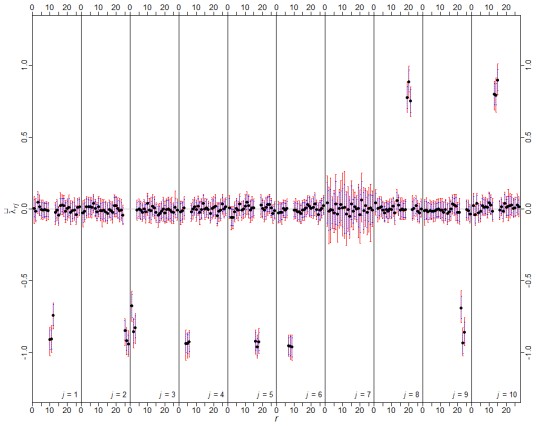
(a) Orthogonal Procrustes with 8 factors



(b) Orthogonal Procrustes with 9 factors



(c) Orthogonal Procrustes with 10 factors



References

Aguilar, O., & West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, *18*(3), 338–357.

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Anderson, T., & Rubin, H. (1956). Statistical inference in. In *Proceedings of the third berkeley symposium on mathematical statistics and probability: Held at the statistical laboratory, university of california, december, 1954, july and august, 1955* (Vol. 1, p. 111).

Aßmann, C., Boysen-Hogrefe, J., & Pape, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics*, *192*(1), 190–206.

Bagby, R. M., Marshall, M. B., & Georgiades, S. (2005). Dimensional personality traits and the prediction of dsm-iv personality disorder symptom counts in a nonclinical sample. *Journal of Personality Disorders*, *19*(1), 53–67.

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 1281–1311.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Bekker, P. A., & ten Berge, J. M. (1997). Generic global indentification in factor analysis. *Linear Algebra and its Applications*, *264*, 255–263.

Berger, J. O. (2013). *Statistical decision theory and bayesian analysis*. Springer Science & Business Media.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., & West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, *7*, 733–742.

Bhattacharya, A., & Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, *98*(2), 291–306.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490.

Bolfarine, H., Carvalho, C. M., Lopes, H. F., & Murray, J. S. (2022). Decoupling shrinkage and selection in gaussian linear factor analysis. *Bayesian Analysis*, *1*(1), 1–23.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3–62.

Bozdogan, H., & Ramirez, D. E. (1987). An expert model selection approach to determine the "best" pattern structure in factor analysis models. In *Multivariate statistical modeling and data analysis: Proceedings of the advanced symposium on multivariate modeling and data analysis may 15–16, 1986* (pp. 35–60).

Burkard, R., & Dell'Amico, S. (2009). Martello assignment problems. *SIAM, Society for Industrial and Applied Mathematics: Philadelphia, PA, USA*.

Cao, Y. (2010). *A bayesian approach to factor analysis via comparing posterior and prior concentration*. Citeseer.

Carneiro, P., Hansen, K. T., & Heckman, J. J. (2003). *Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college.* National Bureau of Economic Research Cambridge, Mass., USA.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, *103*(484), 1438–1456.

Cattell, H. E., & Mead, A. D. (2008). The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, *2*, 135–159.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of econometrics*, *183*(1), 31–57.

Costa, P. T., & McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, *2*(2), 179–198.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, *1928*(1), 13–74.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, *41*(1), 417–440.

Galton, F. (1889). I. co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, *45*(273-279), 135–145.

Gelman, A., & Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 163–185.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*, *4*, 641–649.

Geweke, J., & Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, *9*(2), 557–587.

Geweke, J. F., & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, *75*(369), 133–137.

Goldberg, L. R. (1982). From ace to zombie: Some explorations in the language of personality. *Advances in personality assessment*, *1*, 203–234.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, *4*(1), 26.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161.

Hahn, P. R., & Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, *110*(509), 435–448.

Haslbeck, J., & van Bork, R. (2022). Estimating the number of factors in exploratory factor analysis via out-of-sample prediction errors. *Psychological Methods*.

Heidelberger, P., & Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, *24*(4), 233–245.

Heywood, H. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *134*(824), 486–501.

Huang, A., & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices.

Ihara, M., & Kano, Y. (1995). Identifiability of full, marginal, and conditional factor analysis models. *Statistics & probability letters*, *23*(4), 343–350.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling.

Jennrich, R. I. (1978). Rotational equivalence of factor loading matrices with specified values. *Psychometrika*, *43*(3), 421–426.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

JR, M. (1998). Handbook of matrices h. lutkepohl, john wiley and sons, 1996. *Econometric Theory*, *14*(3), 379–380.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, *20*(1), 141–151.

Karson, S., & O'Dell, J. W. (1976). A guide to the clinical use of the 16 pf.

Kendall, M. G. (1950). Part i: Factor analysis as a statistical technique. *Journal of the Royal Statistical Society. Series B (Methodological)*, *12*(1), 60–73.

Krug, S. E., & Johns, E. F. (1986). A large scale cross-validation of second-order personality structure defined by the 16pf. *Psychological reports*, *59*(2), 683–693.

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *12*(3), 209–229.

Ledermann, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika*, *2*(2), 85–93.

Lee, S.-Y., & Song, X.-Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*, *29*(1), 23–39.

Little, J. D., Murty, K. G., Sweeney, D. W., & Karel, C. (1963). An algorithm for the traveling salesman problem. *Operations research*, *11*(6), 972–989.

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 41–67.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., & West, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian inference for gene expression and proteomics*, *1*(1), 3.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (n.d.).

Marin, J.-M., Mengersen, K., Robert, C. P., Dey, D., & Rao, C. (2005). Bayesian modelling and inference on mixtures of distributions. handbook of statistics 25. *Elsevier-Sciences*.

Martin, J. K., & McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for heywood cases. *Psychometrika*, *40*(4), 505–517.

Martin, R. A. (2001). Humor, laughter, and physical health: methodological issues and research findings. *Psychological bulletin*, *127*(4), 504.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). In-

dividual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of research in personality*, *37*(1), 48–75.

Mavridis, D., & Ntzoufras, I. (2014). Stochastic search item selection for factor analytic models. *British Journal of Mathematical and Statistical Psychology*, *67*(2), 284–303.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, *60*(2), 175–215.

Mcmcpack. (n.d.).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Neuhaus, J. O., & Wrigley, C. (1954). The quartimax method: An analytic approach to orthogonal simple structure 1. *British Journal of Statistical Psychology*, *7*(2), 81–91.

Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology*, *4*(6), 681.

O'Connor, B. P. (2002). A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. *Assessment*, *9*(2), 188–203.

Papastamoulis, P. (2015). label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*.

Papastamoulis, P. (2018). Overfitting bayesian mixtures of factor analyzers with an unknown number of components. *Computational Statistics & Data Analysis*, *124*, 220–234.

Papastamoulis, P. (2020). Clustering multivariate data using factor analytic bayesian mixtures with an unknown number of components. *Statistics and Computing*, *30*(3), 485–506.

Papastamoulis, P., & Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, *19*(2), 313–331.

Papastamoulis, P., & Ntzoufras, I. (2022a). On the identifiability of bayesian factor analytic models. *Statistics and Computing*, *32*(2), 23.

Papastamoulis, P., & Ntzoufras, I. (2022b). On the identifiability of bayesian factor analytic models. *Statistics and Computing*, *32*. doi: 10.1007/

s11222-022-10084-4

Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of personality and social psychology*, *57*(3), 552.

Piatek, R. (2021). Bayesfm: Bayesian inference for factor modeling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFM` (R package version 0.1.5)

Polasek, W. (1997). *Factor analysis and outliers: a bayesian approach*. Citeseer.

Poworoznek, E. (2020). infinitefactor: Bayesian infinite factor models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=infinitefactor` (R package version 1.0)

Poworoznek, E., Ferrari, F., & Dunson, D. (2021). Efficiently resolving rotational ambiguity in bayesian matrix sampling with matching. *arXiv preprint arXiv:2107.13783*.

Raftery, A. E., & Lewis, S. M. (1996). Implementing mcmc. *Markov chain Monte Carlo in practice*, 115–130.

Raymond, B. C. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245–276.

Revuz, D., & Yor, M. (2013). *Continuous martingales and brownian motion* (Vol. 293). Springer Science & Business Media.

Roberts, G. O., & Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, *44*(2), 458–475.

Rodríguez, C. E., & Walker, S. G. (2014). Label switching in bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, *23*(1), 25–45.

Rubin, D. B., & Thayer, D. T. (1982). Em algorithms for ml factor analysis. *Psychometrika*, *47*, 69–76.

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model.

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Spearman. (1904). Objectively determined and measured. *American Journal of Psychology*, *15*(2), 201–292.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Surhone, L. M., Tennoe, M. T., & Henssonow, S. F. (2010). *Openbugs*. Beau Bassin, MUS: Betascript Publishing.

Team, R. C. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological review*, *38*(5), 406.

Thurstone, L. L. (1935). The vectors of mind: Multiple-factor analysis for the isolation of primary traits.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611–622.

Tupes, E. C., & Christal, R. C. (1958). *Stability of personality trait rating factors obtained under diverse conditions* (Vol. 58). Personnel Laboratory, Wright Air Development Center, Air Research and ....

Van Dyk, D. A., & Park, T. (2008). Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, *103*(482), 790–796.

Woodbury, M. A. (1949). The stability of out-input matrices. *Chicago, IL*, *9*, 3–8.

Woodbury, M. A. (1950). *Inverting modified matrices*. Department of Statistics, Princeton University.

Zhang, X., Boscardin, W. J., & Belin, T. R. (2006). Sampling correlation matrices in bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, *15*(4), 880–896.