Model Identification in Bayesian Analysis of Static and Dynamic Factor Models

Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschafts- und Sozialwissenschaften der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Diplom-Volkswirt Markus Pape
aus Bremen

Duisburg, April 2015

Gedruckt mit Genehmigung der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

Dekan:	Professor Dr. Achim Walter
Erstberichterstattender:	Professor Dr. Roman Liesenfeld
Zweitberichterstattende:	Professor Dr. Annekatrin Niebuhr

Tag der Abgabe der Arbeit:	07.01.2015
Tag der mündlichen Prüfung:	27.03.2015

Vorwort

Die vorliegende Arbeit ist während meiner Tätigkeit am Institut für Statistik und Ökonometrie der Christian-Albrechts-Universität zu Kiel und am Seminar für Wirtschafts- und Sozialstatistik der Universität zu Köln entstanden. Kapitel 3, welches das vorgeschlagene Verfahren zur Ex-Post-Identifikation bei der Schätzung statischer und dynamischer Faktormodelle mittels Gibbs Sampling erläutert und dem gängigen Verfahren zur Ex-Ante-Identifikation gegenüberstellt, basiert auf einer gemeinsamen Arbeit mit Herrn Dr. Christian Aßmann und Herrn Dr. Jens Boysen-Hogrefe.

Mein besonderer Dank gilt meinem Doktorvater Herrn Professor Dr. Roman Liesenfeld, der durch seine Anregungen und stets konstruktive Kritik entscheidend zum Gelingen dieser Arbeit beigetragen hat. Dabei habe ich sehr davon profitiert, dass er mir einerseits große Freiräume gestattet hat, um verschiedene Ansätze auszuprobieren und andererseits an wichtigen Stellen entscheidende Hinweise gegeben hat, um die Arbeit in die richtigen Bahnen zu lenken. Danken möchte ich auch Frau Professor Dr. Annekatrin Niebuhr für die Zweitbegutachtung der Arbeit und die inhaltliche Begleitung der arbeitsmarktspezifischen Themen, die in Kapitel 5 untersucht werden.

Meinen Koautoren Herrn Dr. Christian Aßmann und Herrn Dr. Jens Boysen-Hogrefe danke ich besonders für die gute Zusammenarbeit, die vielen ausführlichen und detailreichen Diskussionen und die zahlreichen daraus gewonnenen Erkenntnisse. Frau PD Dr. Sylvia Kaufmann danke ich für die interessanten Diskussionen über die Identifikation von Faktormodellen mit schwach besetzten Ladungsmatrizen und in diesem Zusammenhang insbesondere für wertvolle Hinweise zu Kapitel 4 dieser Arbeit.

Weitere hilfreiche Diskussionen und entscheidende Hinweise verdanke ich außerdem den Teilnehmern verschiedener Seminare, Workshops und Konferenzen, die mit ihrer Expertise zur Entwicklung der Arbeit in der vorliegenden Form beigetragen haben. Hervorheben möchte ich dabei Herrn Dr. Christian Schumacher, Frau Professor Dr. Helga Wagner, Frau Professor Dr. Sylvia Frühwirth-Schnatter, Frau Professor Dr. Bettina Grün, sowie Gutachter und Mitherausgeber des Journal of Econometrics.

Ein großer Dank gebührt außerdem Herrn Professor Dr. Vasyl Golosnoy für die Unterstützung zu Beginn meiner Tätigkeit an seinem Lehrstuhl für Statistik und Ökonometrie an der Ruhr-Universität Bochum, die mir die erforderlichen Freiräume für die Fertigstellung meiner Dissertation ermöglicht hat. Für die freundliche und produktive Arbeits- und Forschungsatmosphäre am Institut für Statistik und Ökonometrie der Christian-Albrechts-Universität zu Kiel und am Seminar für Wirtschafts- und Sozialstatistik der Universität zu

Köln danke ich den Leitern des Instituts bzw. Seminars, insbesondere Herrn Professor Dr. Roman Liesenfeld und Herrn Professor Dr. Helmut Herwartz, sowie Herrn Professor Dr. Uwe Jensen und Herrn Dr. Jan Roestel, und auch meinen langjährigen Kollegen Herrn Dr. Bastian Gribisch, Herrn Jan Vogler und Herrn Jeremias Bekierman, sowie meinen weiteren ehemaligen Kollegen. Entscheidend zum Gelingen dieser Arbeit haben außerdem Herr Dipl.-Informatiker Albrecht Mengel und Herr Julian Schröder beigetragen, denen ich für ihre Unterstützung bei der Durchführung der zahlreichen erforderlichen Berechnungen ausdrücklich danke.

Weiterhin danke ich meinen Freunden und Kommilitonen aus dem PhD-Programm "Quantitative Economics", insbesondere Frau Dr. Laura Birg, Frau Dr. Nadine Heitmann, Frau Dr. Wan-Hsin Liu und Herrn Dr. Matthias Raddant für die gemeinsame Arbeits- und Freizeit.

Schließlich möchte ich einen ganz besonderen Dank an meine Frau Julia richten, die durch ihre liebevolle Unterstützung und ihr Verständnis einen wesentlichen Beitrag zum erfolgreichen Abschluss dieser Arbeit geleistet hat.

Duisburg, den 22. April 2015 Markus Pape

Contents

List of Tables

List of Figures

List of Symbols

This list is not exhaustive, but contains most of the symbols used in this thesis. Symbols used in appendices of single chapters only have generally been omitted from the list. The notation has been adjusted for consistency and double use of the same symbols has been avoided where it might cause misunderstandings. If particular symbols have been used in single chapters or subsections only, this is indicated in the list. If the meaning of particular symbols deviates only in single subsections, this is likewise indicated. The symbols are grouped to make them easier to find in the list.

Sets and Groups

$\mathbb{R}^{M_1 \times M_2 \times \ldots}$	the $M_1 \times M_2 \times \ldots$-dimensional real space
$\mathbb{R}_+$	the real positive numbers
$\mathbb{Z}$	the integers
$O(K)$	the orthogonal group, containing all orthogonal matrices in the $\mathbb{R}^{K \times K}$
$SO(K)$	the special orthogonal group, containing all special orthogonal matrices in the $\mathbb{R}^{K \times K}$
$O(K) \setminus SO(K)$	the set of orthogonal matrices, excluding the set of special orthogonal matrices
$\mathcal{CS}$	constituent set (of angles) (Appendix 3.B)
$\mathcal{K}$	a subset of the axes 1 to K (Chapter 4)
$\mathfrak{K}$	a two-element subset of the axes 1 to K (Chapter 4)

Special Vectors and Matrices

u_i	the i^{th} canonical unit vector of conformable length
I_K	the K-variate identity matrix
$0_{M_1 \times M_2 \times \ldots}$	an $M_1 \times M_2 \times \ldots$-dimensional tensor of zeros
$1_{M_1 \times M_2 \times \ldots}$	an $M_1 \times M_2 \times \ldots$-dimensional tensor of ones

Vector and Matrix Norms

$\|v_i\|$	Euclidean norm of the vector v_i
$\|A\|_F$	Frobenius norm of the matrix $A \in \mathbb{R}^{M_1 \times M_2}$ with $\|A\|_F = \sqrt{\sum_{i=1}^{M_1} \sum_{j=1}^{M_2} a_{i,j}^2}$

Matrix and Other Operators

$\otimes$	the Kronecker product
$\odot$	the Hadamard product (pointwise multiplication)
tr	the trace operator
$\lvert A\rvert$ or $\det(A)$	the determinant of matrix A
$\text{vec}(A)$	vectorization of the matrix A, where all column vectors are stacked into a single vector
$\text{vech}(A)$	half-vectorization of the symmetric matrix A, where the lower triangular part is stacked into a vector
L	the lag operator (Chapter 5)

Probability Density Functions and Probability Mass Functions

$f_N(x\|\mu, \Sigma)$	the pdf of a multivariate normal distribution for the random variable x with mean vector μ and covariance matrix Σ
$f_N(x\|\mu, \sigma^2)$	the pdf of a univariate normal distribution for the random variable x with mean μ and variance σ^2
$f_{IG}(x\|\alpha, \beta)$	the pdf of an inverse gamma distribution for the random variable x with shape and scale parameters α and β
$f_B(x\|\alpha, \beta)$	the pdf of a beta distribution for the random variable x with shape parameters α and β
$\delta_0(x)$	the pdf of a Dirac delta distribution, which has all its probability mass at zero and is therefore a degenerate distribution
π_j	the proportion of mixture component j in a finite mixture distribution (Chapter 2)
$P(\theta)$	distribution on C-group clusterings, which is an $N \times C$ matrix for N observations and C clusters (Section 2.5)
Q	estimated distribution on C-group clusterings (Section 2.5)

Functions and Arguments of Functions

$\mathcal{L}(\cdot)$	likelihood function
$L(\cdot)$	loss function
$\Gamma(\cdot)$	the Gamma function (appears in f_{IG} only)
$\nu(1), \ldots, \nu(K)$	a permutation of the numbers 1 to K
γ	a rotation angle
$\text{mod}(a, b)$	the modulus after a division of a by b
$\lfloor x \rfloor$	the largest integer smaller than or equal to x
$\lceil x \rceil$	the smallest integer larger than or equal to x

q_{xx}	the xx percent quantile of an empirical distribution
α	parameter for the width of an HPDI or HPDE, which is $1-\alpha$
$d_i^{(z)}$	estimated Mahalanobis distance of $\lambda_i^{(z)}$ from the estimated center of its distribution
$\Gamma(\tau)$	autocovariance matrix of order τ
τ	forecast horizon (Chapter 5)
$BF_{0,1}$	Bayes factor comparing models 0 and 1
κ_k	kurtosis of factor k

Matrix Dimensions

T	number of observations for each variable or time series
N	number of variables or time series
P	order of the VAR process governing the dynamic factors
Q	order of the AR processes governing the idiosyncratic components
S	maximum considered lag for loadings of lagged factors

Vectors and Matrices in the Static and Dynamic Factor Model

$Y = (y_1, \ldots, y_T)'$	a $T \times N$ matrix of observable data, where the y_t for $t \in \{1, \ldots, T\}$ are $T \times 1$ vectors
$y_{i,t}$	the i^{th} element of vector y_t
$\tilde{y}_{i,t,k}$	$y_{i,t}$ with the effect of all factors except k removed (Section 4.3)
ϑ	a vector of model parameters
$F = (f_1, \ldots, f_T)'$	a $T \times K$ matrix of factors, where the f_t for $t \in \{1, \ldots, T\}$ are $T \times 1$ vectors
$f_{\cdot,k}$	the k^{th} column vector of F, which is a $T \times 1$ vector
$\{\Phi_p\}_{p=1}^P$	P persistence matrices of dimension $K \times K$ describing the VAR(P) factor process in the dynamic factor model
$\tilde{\Phi} = [\Phi_1', \ldots, \Phi_P']'$	stacked persistence matrices for the factors
$\Phi(L)$	lag polynomial for the VAR(P) factor process (Chapter 5)
ϵ_t	a $T \times 1$ vector of innovations in the factors in the dynamic factor model
Ω	the $K \times K$ covariance of the innovations in the factors ϵ_t (or the factors themselves f_t in the static factor model)
$\Lambda = (\lambda_1, \ldots, \lambda_N)'$	an $N \times K$ matrix of factor loadings, where the λ_i for $i \in \{1, \ldots, N\}$ are the $K \times 1$ vectors of factor loadings on variable i
$\lambda_{\cdot,k}$	the k^{th} column vector of Λ, which is an $N \times 1$ vector
Λ_s	an $N \times K$ matrix of factor loadings with the factors lagged by s periods with $s \in \{0, \ldots, S\}$ in the dynamic factor model

$\bar{\Lambda} = (\Lambda_0', \ldots, \Lambda_S')'$	factor loadings matrix of dimension $(S+1)N \times K$, containing the stacked loadings matrices with loadings of the contemporaneous and lagged factors
$E = (e_1, \ldots, e_T)'$	a $T \times N$ matrix of idiosyncratic components, where the e_t for $t \in \{1, \ldots, T\}$ are $T \times 1$ vectors
$\{\Theta_q\}_{q=1}^{Q}$	Q diagonal persistence matrices of dimension $N \times N$ describing the VAR(Q) process in the idiosyncratic components
$\theta_{q,i,i}$	the i^{th} diagonal element of matrix Θ_q
$\Theta(L)$	lag polynomial for the VAR(Q) process in the idiosyncratic components (Chapter 5)
Θ	an $N \times Q$ matric containing the diagonal elements of all Θ_q for $q \in \{1, \ldots, Q\}$
ξ_t	an $N \times 1$ vector of innovations in the idiosyncratic components
$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$	the $N \times N$ diagonal covariance matrix of the innovations in the idiosyncratic components ξ_t (or the idiosyncratic components themselves e_t if the idiosyncratic components are not autocorrelated)
Δ	an $N \times K$ indicator matrix describing associations between factors and variables for the CFA in the two-step approach

Vectors and Matrices for the Prior Hyperparameters and the Parameters of the Full Conditional Distributions

μ_{λ_i}	prior hyperparameter for the mean of a K-variate normal distribution for the loadings on variable i, generally set to 0_K
Σ_{λ_i}	prior hyperparameter for the covariance of a K-variate normal distribution for the loadings on variable i, generally set to $\underline{c}_i I_K$
$\underline{c}_i$	scaling factor for the prior hyperparameter Σ_{λ_i}
μ_{Λ_s}	prior hyperparameter for the mean of an NK-variate normal distribution for the vectorized loadings matrix vec(Λ_s) (Chapter 3)
Ω_{Λ_s}	prior hyperparameter for the covariance of an NK-variate normal distribution for the vectorized loadings matrix vec(Λ_s) (Chapter 3)
$\underline{\Upsilon}_s$	an $N \times N$ positive diagonal matrix of scaling factors for the prior hyperparameter Ω_{Λ_s} (Chapter 3)
μ_{l_i}	parameter for the mean of a K-variate normal full conditional distribution for the loadings on variable i
Σ_{l_i}	parameter for the covariance of a K-variate normal full conditional distribution for the loadings on variable i

$\mu_{\lambda_{s,i}}$	parameter for the mean of a K-variate normal full conditional distribution for the loadings on variable i with the factors lagged by s periods (Chapter 3)
$\Omega_{\lambda_{s,i}}$	parameter for the covariance of a K-variate normal full conditional distribution for the loadings on variable i with the factors lagged by s periods (Chapter 3)
$\underline{\alpha}_i$	prior hyperparameter of an inverse Gamma distribution for the parameter σ_i^2
$\underline{\beta}_i$	prior hyperparameter of an inverse Gamma distribution for the parameter σ_i^2
a_i	parameter of an inverse Gamma full conditional distribution for the parameter σ_i^2
b_i	parameter of an inverse Gamma full conditional distribution for the parameter σ_i^2
μ_f	prior hyperparameter for the mean of a K-variate normal distribution for the factors in the static factor model, generally set to 0_K
Σ_f	prior hyperparameter for the covariance of a K-variate normal distribution for the factors in the static factor model, generally set to I_K
μ_{f_t}	parameter for the mean of a K-variate normal full conditional distribution for the factors in the static factor model, generally set to 0_K
Σ_{f_t}	parameter for the covariance of a K-variate normal full conditional distribution for the factors in the static factor model, generally set to I_K
$\mu_{\tilde{\Phi}}$	parameter for the mean of a P^2K-variate normal full conditional distribution for the stacked persistence matrices $\tilde{\Phi}$
$\Omega_{\tilde{\Phi}}$	parameter for the covariance of a P^2K-variate normal full conditional distribution for the stacked persistence matrices $\tilde{\Phi}$
$\underline{\zeta}_i$	prior hyperparameter for the mean of a normal distribution for the persistence parameters in the idiosyncratic component for variable i
$\underline{\Psi}_i$	prior hyperparameter for the covariance of a normal distribution for the persistence parameters in the idiosyncratic component for variable i
μ_{θ_i}	parameter for the mean of a Q-variate normal full conditional distribution for the persistence parameters in the idiosyncratic component for variable i
Ω_{θ_i}	parameter for the covariance of a Q-variate normal full conditional distribution for the persistence parameters in the idiosyncratic component for variable i

β_k	prior hyperparameter used to calculate the association probability of all variables with factor k, which is $1-\beta_k$ in the one-layer sparse prior (Section 4.3)
$\beta_{i,k}$	prior hyperparameter for the association probability of variable i with factor k in the two-layer sparse prior (Chapter 4)
ρ_k	prior hyperparameter for the base rate of the association probability of all variables with factor k in the two-layer sparse prior (Chapter 4)
τ_k	prior hyperparameter for the variance of the nonzero loadings on factor k in the two-layer sparse prior (Chapter 4)
$m_{i,k}$	parameter for the mean of the slab part of the full conditional distribution of $\lambda_{i,k}$ in the two-layer sparse prior (Chapter 4)
$M_{i,k}$	parameter for the variance of the slab part of the full conditional distribution of $\lambda_{i,k}$ in the two-layer sparse prior (Chapter 4)
$\underline{c}_0$	prior hyperparameter of an inverse Gamma distribution for the parameter τ_k in the two-layer sparse prior, set to the same value for all $k \in \{1,\ldots,K\}$ (Chapter 4)
$\underline{C}_0$	prior hyperparameter of an inverse Gamma distribution for the parameter τ_k in the two-layer sparse prior, set to the same value for all $k \in \{1,\ldots,K\}$ (Chapter 4)
s_k	prior hyperparameter of a Beta distribution for the association probability $\beta_{i,k}$ in the two-layer sparse prior, set to the same value for all $i \in \{1,\ldots,N\}$ (Chapter 4)
r_k	prior hyperparameter of a Beta distribution for the association probability $\beta_{i,k}$ in the two-layer sparse prior, set to the same value for all $i \in \{1,\ldots,N\}$ (Chapter 4)
w_k	prior hyperparameter of a Beta distribution for the base rate of the association probability ρ_k in the two-layer sparse prior (Chapter 4)
v_k	prior hyperparameter of a Beta distribution for the base rate of the association probability ρ_k in the two-layer sparse prior (Chapter 4)

Matrices Used in Functions and Transformations

D	an orthogonal matrix, sometimes, more generally, an invertible matrix (Section 2.2, 4.2)
B	a reflection matrix, generally of size $K \times K$
P	a permutation matrix, generally of size $K \times K$
Q	a rotation matrix with rotation angle
G	a Givens rotation matrix, performing a rotation around two axes k_1 and k_2

O	an $N \times N$ permutation matrix performing a reordering of the N variables or time series
W	a square matrix of weights
$H(D)$	a matrix to transform the vector of model parameters ϑ based on the transformation of the factors and loadings by the orthogonal matrix D

Other Vectors and Matrices

$M = (\mu_1, \ldots, \mu_n)'$	an $n \times K$ matrix of transposed mean vectors μ_j for $j \in \{1, \ldots, n\}$, where $\mu_j \in \mathbb{R}^K$ (Chapter 2)
$\bar{\Sigma}$	a block-diagonal $nK \times nK$ matrix composed of n $K \times K$ covariance matrices Σ_j for $j \in \{1, \ldots, n\}$ (Chapter 2)
$\tilde{M}$	the parameter of the underlying orthogonally invariant distribution corresponding to the parameter M of the orthogonally mixed distribution (Chapter 2)
$\tilde{\bar{\Sigma}}$	the parameter of the underlying orthogonally invariant distribution corresponding to the parameter $\bar{\Sigma}$ of the orthogonally mixed distribution (Chapter 2)
$\{X_s\}_{s=1}^S$	an orthogonally mixed sample with S elements, where X_s is an $n \times K$ matrix (Chapter 2)
$\{\tilde{X}_s\}_{s=1}^S$	the corresponding sample from the underlying orthogonally invariant distribution (Chapter 2)
$\{D_s\}_{s=1}^S$	the sequence of $K \times K$ orthogonal matrices that the $\tilde{X}_s$ are orthogonally transformed by to obtain the $X_s = \tilde{X}_s D_s$ for $s \in \{1, \ldots, S\}$ (Chapter 2)
$\{E_s\}_{s=1}^S$	the sequence of S $n \times K$ matrices of errors of the underlying orthogonal invariant distribution such that $\tilde{X}_s = \tilde{M} + E_s$ (Chapter 2)

List of Abbreviations

ab	average length of burn-in sequence
ACF	autocorrelation function
AIC	Akaike information criterion
AR	autoregressive
ARIMA	autoregressive integrated moving average
BIC	Bayesian information criterion
CEI	coincident economic indicator
CFA	confirmatory factor analysis
DFM	dynamic factor model
DGP	data generating process
DIC	deviance information criterion
DSGE	dynamic stochastic general equilibrium
EFA	exploratory factor analysis
EM	expectation-maximization
EMU	European Monetary Union
EN	elastic net
ESEM	Econometric Society European Meeting
ETKF	ensemble transform Kalman filter
EU	European Union
FAVAR	factor-augmented vector autoregression (or: autoregressive)
FM2	monetary base
FYFF	federal funds rate
GDP	gross domestic product
HAC	heteroskedasticity and autocorrelation consistent
HPD	highest posterior density
HPDE	highest posterior density ellipsoid
HPDI	highest posterior density interval
ICA	independent component analysis
ICOMP	informational complexity criterion
i.i.d.	independently identically distributed
IP	industrial production
LASSO	least absolute shrinkage and selection operator
LEI	leading economic indicator
LT	lower triangular
MC	Monte Carlo

ML	maximum likelihood
MIMIC	multiple indicator multiple cause
MCMC	Markov Chain Monte Carlo
NAPM	National Association of Purchasing Management
nc	non-converged sequences after 100,000 iterations
NUTS	nomenclature des unités territoriales statistiques (nomenclature of units for territorial statistics)
NUTS-3	territorial units on county level (for Germany)
OP	orthogonal Procrustes
PC	principal components
PCA	principal component analysis
pdf	probability density function
PLT	positive lower triangular
PMCP	NAPM commodity price index
QQ	quantile-quantile
RMSE	root mean-squared error
RMSFE	root mean-squared forecasting error
SEM	structural equation modeling
SVAR	structural vector autoregression (or: autoregressive)
VAR	vector autoregression (or: autoregressive)
WOP	weighted orthogonal Procrustes

Chapter 1

Introduction

1.1 Factor Analysis

Factor analysis serves as a tool for dimensional reduction. Its aim is to find latent variables, the common factors, which drive observable variables, where the dimensional reduction implies that the number of latent factors is much smaller than the number of observable variables. The latent factors are connected to the observable variables via factor loadings, whose values reveal the direction and the strength of the connection. In a factor model, the product of factors and loadings then constitutes the systematic part of the observable variables, whereas an idiosyncratic part remains that is not explained by the common factors. The factors are often determined in such a way that the share of variation in the observable variables that is explained by the factors is maximized, which implies that the factors are mutually orthogonal. On a broader canvas, factor models are closely related to structural equation models, see e.g. Song and Lee (2012), which can be seen as their generalization, and both are themselves members of the much larger family of latent variable models, see e.g. Skrondal and Rabe-Hesketh (2004).

Initially developed in psychometrics as single-factor analysis, see Spearman (1904), and later extended to multiple factor analysis, see Thurstone (1931, 1935), factor analysis has now found a multitude of applications in many different sciences. Model setups have been adjusted according to the requirements of these applications, which often led to the development of new estimation techniques. For instance, in psychometrics, oblique factor models, which allow for correlation among the factors, see e.g. Gorsuch (1983), were introduced to facilitate interpretability of the factors. In economics, on the other hand, the analysis of time series data has led to models with dynamics in the factors and idiosyncratic components, see e.g. Bai and Wang (2012). A rather recent extension of particular interest are sparse factor models, see e.g. Ma and Zhao (2013), developed in biostatistics for gene expression analysis, which are used in cancer research. In a sparse factor model, the loadings matrix contains many zero entries, implying that each variable is only linked to a small subset of the latent factors.

This thesis looks at factor analysis primarily from the perspective of economics and therefore focuses on the model types of particular interest there. Two recent summary papers on factor analysis likewise taking this perspective are Barhoumi et al. (2013) and Lütkepohl (2014).

The former focuses on reviewing the literature, while the latter mainly discusses the technical properties of factor models, taking a wide range of model variations into account.

Whether the latent factors postulated in the factor model setup exist as actual but unobservable variables or are conversely merely a statistical construction exploited for the purpose of dimensional reduction has regularly been debated according to Bartholomew (1984). In many applications, however, the latent factors find a justification in the underlying theory. The first factor model with a single common factor by Spearman (1904) was applied to intelligence test data with the purpose of identifying a "general factor" of intelligence capturing the information from the results of multiple tests. Similarly, factor-analytic approaches to economic time series data often aim at finding business cycle indicators that are responsible for a large share of the dynamics in the series, see e.g. Stock and Watson (1989). In gene expression analysis, the relation between the genetic information of a cell, the genotype, and its appearance, the phenotype, are analyzed.[1] The "interpretation" of the genetic information in a cell consists of multiple steps, the first of which involves transcription factors, which regulate the activation or suppression of the contained information. As each gene is only regulated by a small subset of the transcription factors, the use of a sparse factor model, as proposed by West (2003) and Lucas et al. (2006), is justified in this context. This thesis leans towards the notion of factors as merely a statistical construction, treating the factors as augmented parameters of the model. In Chapter 5, however, the relation between the extracted factors and several business climate indicators is analyzed.

Factor analysis was initially conceived as an exploratory technique that allowed all factors to interact with all variables. Holzinger and Swineford (1937, 1939) later introduced the bi-factor model, in which each variable is connected to a common factor and to an additional group-specific factor. This implies that a large number of the factor loadings are initially set to zero and the factor analysis only serves to determine the remaining free elements of the loadings matrix. This approach was later generalized by Jöreskog (1969) to confirmatory factor analysis (CFA), which allows for nearly arbitrary constraints on the loadings matrix. Conversely, the initial factor analytic approach without such constraints is called exploratory factor analysis (EFA). Sparse factor analysis yields a loadings matrix with many zero entries, but these zero entries are not fixed ex ante, therefore it can also be considered as a type of EFA. This thesis deals almost exclusively with EFA; the term "factor analysis" therefore always refers to EFA in the following, while references to CFA are explicitly stated as such.

Different estimation techniques can be applied for EFA and CFA. Principal components (PC) factor analysis became feasible through Principal Components Analysis (PCA) by Pearson (1901), which predates the "general intelligence factor" hypothesis by Spearman (1904) only by three years. Lawley (1940) introduced the maximum likelihood (ML) estimation approach for factor analysis, later improved and extended by Jöreskog (1967). These improved estimation techniques also made inference in CFA feasible, see Jöreskog (1969, 1979a). Subsequently, the ML approach to factor analysis has typically been pursued using

[1] For the relation of genotype and phenotype in general, see e.g. Johannsen (1911).

the expectation-maximization (EM) algorithm as in Rubin and Thayer (1982). Bayesian factor analysis can be traced back to the Bayesian estimation of the multiple indicator multiple cause (MIMIC) model of Zellner (1970).[2] Moreover, Bayesian factor analysis is explicitly discussed in Press (1972) and developed further by Kaufman and Press (1973) and Press and Shigemasu (1989, 1997), where the marginal posterior distributions are approximated for large samples under specific prior choices. Bayesian inference for CFA is proposed by Lee (1981).

If large sample approximations to the properties of the estimators are to be avoided, high-dimensional integrals have to be solved, which can be achieved by Monte Carlo integration, as shown by Kloek and van Dijk (1978). Accordingly, Geweke and Zhou (1996) propose a Markov Chain Monte Carlo (MCMC) approach based on the Gibbs sampler, see Geman and Geman (1984) or Casella and George (1992). Treating the factors as augmented parameters, see Tanner and Wong (1987), and applying the framework by Chib and Greenberg (1994) to model autocorrelated factors and errors, Otrok and Whiteman (1998) extend the MCMC estimation approach for inference in a dynamic single factor model. MCMC approaches to estimate multi-factor models are proposed by Aguilar and West (2000) and Kose et al. (2003); for a recent comprehensive overview of available estimation procedures for dynamic factor models, see e.g. Bai and Wang (2012). Sparse factor analysis also uses MCMC methods for inference, where hierarchical prior distributions for the loadings are used, see e.g. West (2003), Lucas et al. (2006) and Carvalho et al. (2008).

When the factor model to be estimated is specified, the question about the appropriate number of factors immediately arises. In their introduction of large cross-section approximate dynamic factor models, Stock and Watson (1998) propose a possible criterion to select the number of factors to be used in the model and show that an overestimation of the number of factors does not harm consistency of the parameter estimates, in particular, the idiosyncratic variances. The criterion they use is based on the Bayesian information criterion (BIC) by Schwarz (1978); the heuristic criteria to determine the number of factors introduced in psychometrics e.g. by Kaiser (1960) or Cattell (1966) are not considered. Bai and Ng (2002) provide a set of improved information criteria to select the number of static factors that perform well in simulation studies and provide extension for dynamic factor models in Bai and Ng (2007). Additional criteria to select the number of dynamic factors are proposed e.g. by Amengual and Watson (2007), Hallin and Liska (2007), Jacobs and Otter (2008), Onatski (2010), Kapetanios (2010) and Breitung and Pigorsch (2013).

1.2 Factor Analysis in Economics

A common approach to analyzing a large macroeconomic data set consisting of various series of economic activity is to assume that the comovements in all series can be summarized by one or multiple business cycle factors. As i.a. Diebold and Rudebusch (1996) note, the

[2] This model is technically a single-factor model, where the factor is first estimated from a set of observable covariates and then serves as a covariate in a second regression, see Skrondal and Rabe-Hesketh (2004).

empirical concept of business cycles from the book of Burns and Mitchell (1946) is built on the analysis of comovements of a multitude of economic time series, where turning points in individual time series are dated in order to find the according turning points in the business cycle, while accounting for individual leads and lags in the particular series. Referring to this book, Koopmans (1947) notes that "there is a similarity here with Spearman's psychological hypothesis of a single mental factor common to all abilities" and considers factor analysis a possibly suitable approach for the analysis of macroeconomic data series, at the same time pointing out that identifying a single factor or index "may be a good first approximation", while - unlike in astronomy, which Koopmans uses as a reference point - the analysis of business cycles must "essentially [be treated as] a stochastic process [...] because of the great number of factors at work", in particular the "underlying economic behavior of individuals" that eventually produces the observed macroeconomic time series. This statement can be understood as a call for cautiousness when applying models to macroeconomic data and attempting to obtain estimates for the model parameters therefrom.[3] Factor analysis seems an especially adequate tool for the purpose of analyzing large macroeconomic data sets, following the advice of Koopmans (1947) not to dismiss the theoretical macroeconomic foundations altogether, and at the same time not "pretending to have too much a priori economic theory", see Sargent and Sims (1977).

Gathering stylized facts about business cycles, Lucas (1977) provides a detailed description of the comovements of economic time series, which are overall present in series across broadly defined sectors, more pronounced in output series for durable goods and business profits, and less pronounced in prices and output figures of commodities. The vector-autoregressive (VAR) model introduced by Sims (1980) allows to incorporate the comovements in multiple time series to some extent, though due to the fact that in an unconstrained model, the number of parameters grows linearly in the number of lags included in the model, and quadratically in the number of time series that enter the model, macroeconomic time series generally do not provide sufficiently many observations for the model to be identified. This feature is generally referred to as the "curse of dimensionality", a term coined by Bellman (1961) and used in the context of VAR models e.g. by Hendry and Doornik (1994), who argue in favor

[3] The statement regarding structural variability of economic models predates by almost three decades the famous critique by Lucas (1976) that macroeconomic models should anticipate agents' responses to policy changes rather than assuming that model parameters estimated from historical data stay put, which is a reasonable assumption in the aforementioned biostatistics models. The subsequently proposed macroeconomic models take the Lucas critique into account. In turn, a very successful class of micro-founded macroeconomic models emerged: the dynamic general stochastic equilibrium (DSGE) models, see e.g. Kydland and Prescott (1982) and Rotemberg and Woodford (1997).
Two issues of DSGE models turned out to be the assumption of homogenous agents and their rational expectations, see Tinbergen (1932) or Muth (1961). Macroeconomic models not taking heterogeneous agents into account are criticized e.g. by Forni and Lippi (1997), who assert that the models can be saved by introducing according heterogeneity on the micro-level. Similarly, problems with the rational expectations hypothesis have been known for a long time, see e.g. Shiller (1978), but widely neglected until the recent financial crisis. Extended DSGE models, however, account for bounded rationality by incorporating findings from behavioral economics, see e.g. De Grauwe (2010).
So despite the challenges that the foundations of macroeconomics are regularly exposed to, it does not seem adequate to give up on the analysis of macroeconomic data altogether, but rather apply empirical methods that leave sufficient leeway to discover appropriate models, keeping in mind that "all models are wrong, but some are useful", see Box and Draper (1987).

of "parsimonious VAR" models, nowadays mostly referred to as constrained, or structural VAR (SVAR) models. How the model structure is found, however, remains the choice of the practitioner, who can apply model selection procedures like the one proposed by Krolzig (2003), or Bayesian techniques introducing prior beliefs e.g. using the approach of Banbura et al. (2010). Factor models are yet a different way to deal with large-dimensional data sets, and may even "turn the curse of dimensionality into a blessing", see Stock and Watson (2012), being able to use information from a large number of series to estimate the factors, see Stock and Watson (1998).

Cattell (1952) suggests that factor analysis could be applied to repeated observations on the same subject, hence, to time series data potentially subject to serial correlation. Critically analyzing this approach, Anderson (1963) discourages the use of factor analysis for time series data, pointing out e.g. that the assumption of independent error terms may not hold in time series data, and that estimating a static factor model for time series data ignores the possibility of serial dependence in the factors.[4] Contemporaneously, Brillinger (1964) devises his frequency-domain principal components approach, summarized in Brillinger (1981), so the subsequently developed dynamic factor models for macroeconomic data analyze the data in the frequency domain to extract an unobservable index, or latent factor, see Geweke (1977) and Sargent and Sims (1977). These models focus on rather small cross-sections, often only slightly beyond the scope of unconstrained VAR models with respect to the number of parameters. The generalized dynamic factor model brought forth by Forni et al. (2000) is a generalization of the model by Geweke (1977) and Sargent and Sims (1977) in such a way that the idiosyncratic components are no longer constrained to orthogonality. The underlying theory extends the results of Brillinger (1981) and is analyzed e.g. by Forni and Lippi (2001), Forni et al. (2004) and Forni et al. (2005), who show that dynamic principal components can be used to consistently estimate the dynamic factors.

A dynamic model for the time domain, conversely, is proposed in Sargent and Sims (1977) as an observable index model, in fact a reduced-form SVAR model. A model with a single unobservable factor is subsequently proposed by Engle and Watson (1981), who introduce state-space methodology to estimate the unobservable index[5] in the time domain and enhance their Kalman filtering approach by the EM algorithm of Dempster et al. (1977) in their follow-up paper, see Watson and Engle (1983).[6] An extension for multiple factors, or indices,

[4] Recent inference techniques in dynamic factor analysis, such as Doz et al. (2011), however, account for the dynamic features of the factors by applying a two-step estimation approach.

[5] Lütkepohl (2014) calls the factor and index models "closely related". The terms are generally used interchangeably, where single-factor models are often called index models, particularly in finance applications.

[6] Rubin and Thayer (1982) also use the EM algorithm to estimate a factor model, albeit their psychometrics application makes use of a static model. Presumably unaware of the papers by by Sargent and Sims (1977), Engle and Watson (1981) and Watson and Engle (1983), Molenaar (1985) discusses the use of the results by Brillinger (1964, 1981) to estimate latent factors for sufficiently long time series in the frequency domain, and likewise propose the use of the Kalman filter. Unlike Engle and Watson (1981) however, who rely on the work of Mehra (1974), which introduces state space methodology known from engineering to economics and finance, Molenaar (1985) directly draws from the engineering text of Jazwinski (1970). It appears that, while e.g. Anderson and Rubin (1956) and Goldberger (1972) are still able to provide a rather complete overview of the advances in the study of factor models and structural equation models, the subsequent findings in

is proposed by Reinsel (1983), who points out the relation to the reduced rank regression of Izenman (1975).[7] On the side of macroeconomic applications, the important question of how to construct relevant indices, such as coincident economic indicators (CEI), leading economic indicators (LEI) and recession indicators, is discussed by Stock and Watson (1989), who propose to select up to 11 out of 280 available series to estimate the CEI and LEI, respectively. Hence due to the limitations of the model, information in the remaining series is not used. In order to make use of additional series, Quah and Sargent (1993) propose an approach to estimate dynamic indices from random fields, where the number of cross-sections and the number of observations is comparable. Using 60 time series, this is the first large-dimensional dynamic factor model in the time domain.

While the initially proposed factor models from psychometrics generally assume no cross-correlation in the error terms, hence postulating that all comovements are explained by the factors alone, the aforementioned class of dynamic factor models estimated in the time domain allows for a general error covariance matrix in order to account for cross-correlated error terms. A third approach is proposed by Chamberlain and Rothschild (1983), who denote the classic factor model approach as "exact factor models" and introduce the complementary "approximate factor models", where weak cross-correlation in the idiosyncratic components exists, but is not accounted for in the estimation procedure, which uses principal components. The error induced by this misspecification is shown to disappear as the number of cross-sections and the number of observations gets large. Connor and Korajczyk (1986, 1988, 1993) extend their approach, showing that the estimation is likewise consistent if sequential asymptotics instead of joint asymptotics are considered. The approximate static factor model design is then augmented by Kalman filtering in the paper of Stock and Watson (1998), which introduces approximate dynamic factor analysis for large cross sections, analyzing a data set consisting of 224 time series. In a follow-up paper, Stock and Watson (1999) demonstrate how a generalized Phillips curve using accordingly estimated indices provides improved forecasts. Correlated factors arise if the factors are jointly modeled as a VAR process with nonzero off-diagonal elements in the persistence matrices. To obtain a model suitable for forecasting multiple series while using information from a large number of additional series, Bernanke et al. (2005) introduce the factor-augmented VAR (FAVAR) model, which is essentially a VAR model with observable and unobservable factors, estimating a macroeconomic model with one observable and three unobservable factors in their application.[8] The aforementioned DSGE models are paired with factor models by Boivin and Giannoni (2006). The estimation

either psychometrics or econometrics go largely unnoticed by the respective counterparts. Goldberger (1972) remarks that the technical issues occurring with causal analysis are often the same across a variety of fields that uses these models. This idea has recently been brought back to mind by Heckman and Pinto (2013), not only taking methodological findings into account, but also incorporating the qualitative psychological results in the model setup, see e.g. Heckman et al. (2014). Conversely, advances in factor analysis from the field of biostatistics have quickly found applications in econometrics, see e.g. Kaufmann and Schumacher (2012, 2013).

[7] Chan et al. (2013) pick up this relation to deal with the ordering problem in factor analysis that is central to this thesis.

[8] The application in Chapter 3 uses the same data set and finds that if only latent factors are estimated, the additional unobservable factor has a correlation of 0.998 with the observable factor from Bernanke et al. (2005).

approach for the dynamic factor model in the time domain most popular among practitioners in macroeconomics nowadays is the iterative principal components-based approach by Doz et al. (2011) and its quasi-maximum likelihood variant from Doz et al. (2012).

This thesis deals both with static and dynamic factor models, where the factors in the static models are always assumed to be orthogonal, and the factors in the dynamic models are always allowed to be correlated through an unconstrained VAR process, as in Bai and Wang (2012). All factor models are analyzed in the time domain.

1.3 Object of Investigation

Due to the fact that the factors are latent and the loadings are unknown parameters, the factor model suffers from indeterminacies even if a clear distinction between the systematic and the idiosyncratic part is possible. This was already recognized by Thurstone (1935), who states that "In order that a unique solution of [the factors] may be found for any given [correlation matrix], it will therefore be necessary to impose further restrictions on the solution." It depends on the applied estimation technique whether these indeterminacies affect the estimation process or are merely of interest when interpreting the results. If the estimation process is affected, a small number of identifying constraints must be imposed to obtain parameter estimates.

In PC factor analysis, the factors are constructed from principal components, which are uniquely identified if the corresponding eigenvalues of the sample covariance or sample correlation matrix are unique. This is generally the case for the principal components corresponding to the largest eigenvalues, so no further constraints are necessary for identification during the estimation process. After obtaining the parameter estimates, however, it is possible to improve interpretability by transforming them, e.g. creating a *simple structure* in the loadings matrix, see Thurstone (1935). Several rotation techniques, orthogonal and oblique, have been proposed for that purpose, such as the Varimax rotation, see Kaiser (1958).

In ML factor analysis, on the other hand, constraints are necessary to ensure that the likelihood has only a single global mode. The identification problem incurred otherwise is due to the possibility of transforming the latent factors by an invertible matrix, and the loadings matrix by its inverse. The constraints must therefore rule out that such a transformation of the factors and loadings exists, and can be chosen in different ways, see e.g. Millsap (2001). One approach to deal with the identification problem is to split it up into a *scaling problem* and a *rotation problem*, see e.g. Anderson and Rubin (1956), which are then separately solved. The scaling problem is solved by fixing the factor covariance matrix, and the rotation problem is solved by imposing a positive lower triangular (PLT) structure onto the loadings matrix, see e.g. Muirhead (1982). This thesis shows that many constraints equivalent to the PLT structure can be found that obtain the same unique maximum of the likelihood. The values

of the unconstrained parameters in the maximum of the likelihood change accordingly, and the shape of the likelihood depends on which set of constraints is chosen.

Contemporary Bayesian factor analysis typically uses the MCMC method of Gibbs sampling, see e.g. Geman and Geman (1984) and Casella and George (1992), to obtain samples from the posterior distributions of the model parameters and factors. Assuming a quadratic loss function, the Bayes risk is minimized if the posterior mean serves as the Bayes estimator, see e.g. Berger (1985). The sample means then serve as Monte Carlo estimates of the posterior means. Bayesian factor analysis with Gibbs sampling is a very versatile simulation-based approach, allowing for inference in models with dynamics in the factors and loadings. In such setups, the Gibbs sampler for state-space models proposed by Carter and Kohn (1994) can be applied.

Model identification in this Bayesian framework can be achieved by an appropriate choice of prior distributions. In a model with static factors, the scaling problem can be solved accordingly to ML factor analysis, by fixing the factor covariance matrix to the identity matrix. In a model with dynamic factors, the scaling identification is achieved not by fixing the covariance matrix of the factors, but setting the covariance matrix of the innovations in the factors to the identity matrix. The dynamic factors are jointly modeled as a VAR process without constraints, which hence allows them to be correlated. In both models, with static and dynamic factors, the rotation problem therefore remains unsolved. To solve it, prior information about the loadings parameters would have to be used. In a purely exploratory analysis, however, such information is usually not available. To obtain a unimodal posterior distribution, an approach similar to imposing the PLT constraints in ML factor analysis can be chosen. This approach, proposed by Geweke and Zhou (1996), uses informative prior distributions for the upper triangular elements of the loadings matrix, namely Dirac delta priors for the elements above the diagonal - which fixes these elements to zero - and normal priors truncated below at zero for the elements on the diagonal. As a consequence, every sampled loadings matrix satisfies the PLT constraints. This identification approach has become the standard way to ensure a unique identification of the factors and loadings in Bayesian factor analysis by Gibbs sampling, both for exploratory factor analysis, see e.g. Bai and Wang (2012) and for sparse factor analysis, see Carvalho et al. (2008).[9]

If the observable variables are reordered and the PLT constraints are imposed on the loadings matrix of the model for the reordered data, the model is identified in a fashion equivalent to the above, with the exception that the resulting factors and loadings are orthogonal transformations of the factors and loadings of the initial model. The same holds if the factors are reordered and the PLT constraints are imposed on the accordingly adjusted loadings matrix. It has been observed, however, that inference results vary substantially with the ordering of the variables and the estimates are not merely orthogonal transformations of each

[9] A similar approach, which solves the scaling problem and the rotation problem in a single step and avoids the use of truncated prior distributions is proposed by Aguilar and West (2000), who leave the factor variances unconstrained and instead fix the diagonal elements of the loadings matrix to one.

other. This is known as the ordering problem and has been discussed by Lopes and West (2004), Carvalho et al. (2008) and Frühwirth-Schnatter and Lopes (2012).

In the Gibbs sampler for the model with dynamic factors applied in this thesis, the standard Kalman filter is replaced by the faster and more precise ensemble transform Kalman filter of Tippett et al. (2003). The sampling approach thus closely follows Bai and Wang (2012). This thesis investigates the behavior of the Gibbs sampler and the shape of the posterior distributions if the PLT constraints are imposed during the sampling process and thus analyzes the ordering problem.

As the ordering problem is apparently caused by the PLT constraints, it seems adequate to design a sampler which works without such constraints. A Gibbs sampler applied to a model where the scaling problem is solved by the appropriate choice of hyperparameters, but where the rotation problem remains unsolved is described and analyzed in this thesis. Obviously, the output of this sampler is unfit for inference, since the underlying model is not identified. To solve this problem, a postprocessing procedure is proposed, which is called Weighted Orthogonal Procrustes (WOP) approach.

If the output of the unconstrained Gibbs sampler postprocessed with WOP does not rely on ex-ante constraints to solve the rotation problem, and moreover, the prior distributions can be chosen such that the posterior distributions are fully orthogonally invariant, orthogonal transformations of the postprocessed Gibbs output are admissible in a way similar to the well-known orthogonal transformations of the estimates obtained from PC factor analysis that are used to establish a simple structure in the loadings. As the samples from the posterior distributions are much more informative than the results of PC factor analysis, however, they can be used to calculate highest posterior density (HPD) intervals to determine the number of factors, to distinguish between relevant and irrelevant variables and to find a sparse model representation. This aspect is also investigated in this thesis.

Eventually, these techniques are applied to identify the common factors in the dynamics of regional labor markets, where regional labor market data for 402 German counties is analyzed, taking both a full and a sparse factor model into account.

1.4 Outline

Chapter 2 outlines the properties of the proposed Gibbs sampling approach, called the unconstrained Gibbs sampler. Due to the fact that no constraints are imposed on the loadings matrix, the rotation problem remains unsolved and the corresponding indeterminacy about the factors and loadings remains. Therefore, arbitrary orthogonal transformations of the sampled factors and loadings matrices occur during the sampling process. The posterior distributions of the factors and loadings resulting from the indeterminacy are called orthogonal mixture distributions, and the sampler is called orthogonally mixing. To illustrate the properties of orthogonal mixture distributions and to motivate how samples from such distributions can be

further processed, properties of orthogonal matrices are discussed. Next, the mixing in the sampling process is analyzed for particular subgroups of orthogonal matrices.

One of these subgroups are permutation matrices, which, if occurring in the sampling process, cause label switching. This is a phenomenon well known from Gibbs samplers for Markov switching and mixture models, see e.g. Frühwirth-Schnatter (2006). To make inference feasible in these models, the label switching must be suppressed by effectively chosen constraints. Unfortunately, such labeling constraints have often been found to be "ineffective in removing the symmetry in the posterior distribution. As a result, problems with label switching may remain" even after imposing them, "if the constraint is not carefully chosen.", see Stephens (2000). Carefully choosing the constraints, however, is usually not possible ex ante, as reliable information to base the choice of constraints on is not at hand. Therefore, Stephens (2000) proposes to run a sampler that is unconstrained with respect to labeling and therefore subject to random label switching, a concept that is extended by Frühwirth-Schnatter (2001), who proposes to include an additional random label switching step to improve the sampler's mixing behavior. In a subsequent step, an iterative relabeling algorithm is run on the obtained sequence from the unconstrained sampler, see e.g. Stephens (2000) and Jasra et al. (2005).

A similar phenomenon is sign switching, which is caused by another subgroup of orthogonal matrices, reflection matrices. Sign switching occurs e.g. in Bayesian CFA. Therefore, instead of potentially choosing ineffective sign constraints, such constraints can be omitted, allowing for sign switching in the sampling process and afterwards iteratively readjusting the column signs for each draw. Such an approach along the lines of Stephens (2000) has been suggested by Erosheva and Curtis (2013) and is also applicable for Bayesian EFA in the case of a single factor.

Orthogonal mixing in a general sense, as occurring in multi-factor Bayesian EFA, however, is not restricted to label switching and sign switching, i.e. it does not only involve permutation and reflection matrices, but also rotation matrices. The aforementioned algorithms for label and sign switching therefore do not suffice to postprocess the output of an orthogonally mixing sampler. Thus, the orthogonal Procrustes (OP) algorithm is introduced, which relies on orthogonal Procrustes transformations, see e.g. Schönemann (1966) or Golub and van Loan (2013), of the draws in the unconstrained Gibbs output. The procedure works similarly to the relabeling proposed by Stephens (2000). To account for known variation in the loadings vectors per variable, the approach is augmented by a weighting scheme, hence performing weighted orthogonal Procrustes transformations, see e.g. Lissitz et al. (1976) or Koschat and Swayne (1991). This postprocessing procedure is therefore called weighted orthogonal Procrustes (WOP) procedure.[10]

Two simulation studies show that the WOP approach is able to remove orthogonal mixing from samples following different multivariate elliptical distributions, investigate its convergence properties and the tail properties of the recovered empirical distributions. Applying the

[10] The weights are chosen such that the determinant of the covariance matrix of each loadings vector is set to one.

unconstrained sampler, the PLT constrained sampler and the unconstrained sampler with WOP postprocessing to a small static factor model setup, each sampler's behavior is investigated under three different orderings of the data. One of the orderings corresponds to a "effectively chosen" set of constraints in the PLT constrained sampler, which then behaves almost identically to the unconstrained sampler with WOP postprocessing. Under the other two orderings, the corresponding constraints are unable to fully suppress orthogonal mixing, analogously to the observations of Stephens (2000) with respect to label switching. Conversely, the unconstrained sampler successfully removes all orthogonal mixing induced by the unconstrained sampler.

Chapter 3 discusses Bayesian estimation of static factor models in more detail and extends the analysis to dynamic factor models. It looks at the implications of the rotation problem, which is described e.g. in Anderson and Rubin (1956). In Bayesian factor analysis, the rotation problem is dealt with by choosing priors that constrain the parameter space, see e.g. Bai and Wang (2012). Alternatively, it can remain unsolved while sampling, allowing for the generation of an orthogonally mixed sample, which is afterwards postprocessed under a quadratic loss function, using the WOP approach and an additional numerical optimization. A simulation study compares the proposed approach to the commonly used ex-ante model identification imposing positive lower triangularity (PLT) constraints on the loadings matrix, as introduced for Bayesian factor analysis by Geweke and Zhou (1996). Issues arising in the context of the PLT approach, such as order-dependence and multimodality, are discussed. Moreover, quantities unaffected by the rotation problem are analyzed. In a subsequent simulation study, it is shown that the WOP approach provides a remedy to these issues. An empirical study, applying the WOP approach to a dynamic factor model setting, analyzing a data set of 120 macroeconomic time series from a study by Bernanke et al. (2005), confirms the findings from the simulation exercise.[11]

Chapter 4 discusses sparse factor models, see e.g. West (2003), originally introduced in biostatistics, but increasingly popular in the analysis of large macroeconomic data sets, see e.g. Kaufmann and Schumacher (2012, 2013). The purpose of sparse factor analysis is to find a parsimonious structure in the loadings matrix. Sparse factor analysis is therefore closely related to confirmatory factor analysis, with the important difference that in confirmatory factor analysis, the sparse structure is assumed to be known, and the estimation is conducted conditional on this structure. The question of model identification is more complicated in confirmatory factor analysis than in exploratory factor analysis and has been discussed extensively for ML confirmatory factor analysis, where several authors have pointed out cases where models are not identified, see e.g. Dunn (1973) and Jennrich (1978). If conditions for identification given in Bekker (1986) are satisfied, however, the model likelihood has a unique global maximum. A second issue is the multimodality of the model likelihood. Its surface may make it extremely complicated to find the unique maximum among a multitude of local

[11] An empirical study applying the WOP approach to a static factor model setting, analyzing a data set of ten equity indices similar to the one used in Geweke and Zhou (1996), likewise shows the advantages of the WOP approach. The study is not part of this thesis, but can be found in Aßmann et al. (2012).

maxima, see Millsap (2001) and Loken (2005). This issue affects Bayesian factor analysis in a similar way, obstructing the Gibbs sampler from exploring the full posterior distribution. As a result of these numerical issues, the sampler may display spurious convergence. To prevent overlooking such multimodality in the posterior distribution, the WOP approach can be applied to find highest posterior density (HPD) intervals for all entries of the loadings matrix. Just as the WOP output can be arbitrarily orthogonally transformed, the HPD intervals can be transformed in the same way. This allows to explore various sparse patterns in the loadings matrix, where entries of the loadings matrix are set to zero if the corresponding HPD intervals contain the zero. The same approach also allows to determine the number of factors in the model. To that end, an orthogonal transformation must be found such that the HPD intervals for all loadings on a factor, except for two or less, contain the zero. In this case, the corresponding factor is no longer identified and can be removed. Accordingly, if the multivariate HPD interval for the loadings on one variable contains the origin, this variable has exclusively zero loadings and can be removed from the sample.

Using data sets simulated according to a data generating process (DGP) from Lopes and West (2004) and Frühwirth-Schnatter and Lopes (2012), the procedure is tested and found to perform very well on the tasks of finding the correct number of factors, eliminating irrelevant variables and recovering the sparse structure used to simulate the data. Next, the procedure is tested on a well-known psychometric data set from Holzinger and Swineford (1939), which contains 301 observations of high school students, who each performed 24 different tests. The data set is assumed to allow for the extraction of a general intelligence factor and several group factors, where the general factor is related to all the tests and the group factors are related only to specific subsets of tests. While Holzinger and Swineford (1939) predefine the loadings structure and perform a confirmatory factor analysis, the approach taken here is purely data-based. The results, however, are very similar: The general factor is clearly identified, several of the group factors are also found.

Chapter 5 applies the WOP approach and its extension for sparse factor analysis to a data set of unemployment figures for all 402 counties of Germany, which are observed over 82 months from January 2007 until October 2013. To ensure stationarity, the data are first transformed into growth rates, hence 81 observations over time remain. As business cycle data are usually unavailable at this temporal and spatial disaggregation, exposure of the counties to the overall business cycle can be approximated by using employment figures, see also Hamilton and Owyang (2012). Similarly, if an according sparse loadings structure is identified, a decomposition into national, regional and idiosyncratic contributions to the unemployment per county can be found, similar to Kose et al. (2003). Initially, some issues of model identification in dynamic factor analysis are discussed. The sampling approach is briefly introduced, and two model selection criteria are described. When applied to different parameterizations, the model selection criteria favor parsimonious lag structures in the factors and errors, but differ with respect to the number of factors. A model with seven factors, one lag in the factor process and no autocorrelation in the idiosyncratic error terms is chosen, which is supported by further model diagnostics. The parameter estimates from the according

dynamic factor model estimated using the WOP approach afterwards undergo orthogonal transformations to find interpretable factors. After an appropriate orthogonal transformation, the dynamic factors obtained from the WOP approach are almost perfectly correlated with the first principal components. Another transformation allows to find a high negative correlation of the first factor with the ifo business climate indicator. In the next step, a sparse model is estimated. Depending on the width of the HPD intervals used to identify a parsimonious loadings structure, the number of factors can be reduced to six or four, of which the first three are very similar and have a substantial number of nonzero loadings. The first factor loads on most counties and is again found to have a strong correlation with the ifo business climate indicator. The second and third factor have nonzero loadings in particular in the north eastern parts of the country and either replace or complement the first factor there. Thus these factors can be understood as local cycles in addition to the overall business cycle, or in its place. Eventually, the forecasting performance of the full and sparse factor models is compared to a simple AR(1) model, where both factor model approaches are found to perform worse at different forecast horizons. The sparse factor model, however, shows a better forecasting performance than the full one when the forecasting horizon is extended.

Chapter 6 summarizes the main results of the thesis.

Chapter 2

Orthogonal Mixtures

2.1 Introduction

This chapter introduces the concept of orthogonal mixture distributions, or orthogonal mixtures. Orthogonal mixtures are finite or infinite mixture distributions whose mixing is governed by orthogonal matrices, which can be permutation, rotation or reflection matrices, or products thereof. The properties of the orthogonal mixtures depend on the type of orthogonal matrices involved. They are of interest in exploratory Bayesian analysis of static and dynamic factor models.

The purpose of factor analysis is to find a small number of common, but unobservable components, the latent factors, that provide a structure for the variation in a large number of variables. The latent factors are connected to the observable data via factor loadings, so the product of factors and loadings forms the systematic part of the factor model, and an idiosyncratic component remains in each variable. With the factors latent and the loadings unknown, both are only jointly identified unless further constraints are imposed to get rid of this indeterminacy. This identification problem in the factor model can be split up into a scaling problem, which is easy to solve, and a rotation problem, which is much harder to solve.

In Bayesian analysis of factor models, estimates for the parameters and the factors are obtained from their posterior distributions. The Bayes estimator is then the estimator that minimizes the Bayes risk, which is the same as the Bayesian expected loss, conditional on a specified loss function, see Berger (1985). If a quadratic loss function is specified, the Bayes estimator is the mean of the posterior distribution. The posterior distribution, however, is often not analytically tractable and hence, its mean cannot be analytically derived. In this case, Monte Carlo (MC) integration, see Kloek and van Dijk (1978) can be used. The particular method to achieve this in Bayesian factor analysis is the Markov Chain Monte Carlo (MCMC) method of Gibbs sampling[1], in which samples from the posterior distribution of the model parameters and factors are simulated by iteratively drawing from their full conditional distributions, see e.g. Geman and Geman (1984) and Casella and George (1992).

[1]Some samplers used in Bayesian factor analysis are not Gibbs samplers in the strict sense, as they include additional Metropolis-Hastings steps.

The reason to leave the factor model partially unidentified lies in the fact that the necessary identification constraints can be chosen in various ways, each resulting in an exactly identified model in a statistical sense. Nonetheless, the choice of constraints is crucial for the sampler's numerical properties and hence for the inference results. A "bad choice" may lead to a poor ability of the sampler to numerically handle the problem that should be fixed by the constraints - in this case, the rotation problem. Unfortunately, the decision about which set of constraints is chosen has to be made ex ante, when usually very little information is available that a "good choice" of constraints could be based on. This difficulty with the choice of constraints and its consequences on inference results has been referred to as the *ordering problem* by Lopes and West (2004), Carvalho et al. (2008) and Frühwirth-Schnatter and Lopes (2012).

A similar observation has been made by Stephens (2000) and Frühwirth-Schnatter (2001) in Bayesian analysis of Markov switching and mixture models. In these models, the state-specific or component-specific parameters may change their positions at arbitrary points in the Gibbs sampling process, a phenomenon denoted as *label switching*. For instance, in the case of a univariate mixture with two components, in one iteration, the Gibbs sampler produces a draw of $(\mu_1,\ \mu_2)'$, but in the next iteration, it produces a draw of $(\mu_2,\ \mu_1)'$. To prevent this, it is possible to impose identifying constraints that fix the labels ex ante, such as the condition $\mu_1 < \mu_2$. Stephens (2000) and Frühwirth-Schnatter (2001), however, find that fixing labels accordingly often fails to successfully discriminate between mixture components or states. They therefore advocate sampling under omission of labeling constraints. As a consequence, the label switching is not suppressed in the sampling process and is still present in the Gibbs output, see e.g. Jasra et al. (2005). In a subsequent step, a postprocessing algorithm is used to relabel the states or components, after which inference on the parameters is possible.

In Bayesian factor analysis, a similar approach may be applicable to avoid the ordering problem. The corresponding unconstrained Gibbs sampler would abstain from imposing ex-ante constraints to solve the rotation problem and a postprocessing procedure similar to the relabeling approach of Stephens (2000) would be required to fix the indeterminacy remaining in the Gibbs sequence. To find such a postprocessing procedure, however, the properties of the output of the unconstrained sampler have to be analyzed, which is done in this chapter.

The chapter proceeds as follows: Section 2.2 illustrates the model indeterminacy and in particular the rotation issue for the latent factor model with static factors and describes a way of solving these issues in maximum likelihood factor analysis. Section 2.3 describes the Gibbs sampler for this factor model and discusses how the model indeterminacy is usually handled by imposing constraints. It also points out the pitfalls of imposing these constraints, relating them to Bayesian analysis of Markov switching and mixture models. To avoid these issues, a Gibbs sampler without these constraints is proposed. This sampler generates samples from orthogonal mixture distributions. Section 2.4 discusses the concept of orthogonal mixtures in more detail, Section 2.5 illustrates the relationship between label switching and orthogonal mixtures, Section 2.6 introduces the approach to remove orthogonal mixing from a sample, Section 2.7 reports the results of a simulation study in which artificial data from different orthogonal mixture distributions is created and the proposed algorithm is used to unmix

the data and estimate the mean. Section 2.8 extends the simulation study and analyzes the empirical distributions of the unmixed data in detail. In Section 2.9, the behavior of the constrained and unconstrained Gibbs samplers from Section 2.3 and the unconstrained Gibbs sampler combined with the postprocessing approach from Section 2.6 are compared for a small artificially created data set. Section 2.10 concludes.

2.2 Indeterminacy and the Rotation Problem

To illustrate the sources of indeterminacy in the factor model and, in particular, the rotation problem, consider a static factor model along the lines of Anderson and Rubin (1956). Denote the observable data $Y = (y_1, \ldots, y_T)'$, where $y_t \in \mathbb{R}^N$, for $t \in \{1, \ldots, T\}$, so N denotes the number of variables, T denotes the number of observations per variable, and Y is a $T \times N$ matrix. Accordingly, denote the latent factors $F = (f_1, \ldots, f_T)'$, where $f_t \in \mathbb{R}^K$ for $t \in \{1, \ldots, T\}$, so K denotes the number of latent factors, with $K << N$, and F is a $T \times K$ matrix. The factor loadings are then $\Lambda = (\lambda_1, \ldots, \lambda_N)'$, where $\lambda_i \in \mathbb{R}^K$ for $i \in \{1, \ldots, N\}$, so Λ is an $N \times K$ matrix. Eventually, the idiosyncratic components are $E = (e_1, \ldots, e_T)'$, where $e_t \in \mathbb{R}^N$ for $t \in \{1, \ldots, T\}$, so E is a $T \times N$ matrix.[2]

In vector form, the model can be written as

$$y_t = \Lambda f_t + e_t \qquad \text{for all} \quad t \in \{1, \ldots, T\}, \tag{2.1}$$

where the factors and errors have zero mean, so

$$E[f_t] = 0 \quad \text{and} \quad E[e_t] = 0 \qquad \text{for all} \quad t \in \{1, \ldots, T\}, \tag{2.2}$$

and are uncorrelated with each other, i.e.

$$E\left[\begin{pmatrix} f_t \\ e_t \end{pmatrix}\begin{pmatrix} f_t \\ e_t \end{pmatrix}'\right] = \begin{pmatrix} \Omega & 0 \\ 0 & \Sigma \end{pmatrix} \qquad \text{for all} \quad t \in \{1, \ldots, T\}. \tag{2.3}$$

In matrix form, the model can be written as

$$Y = F\Lambda' + E. \tag{2.4}$$

2.2.1 Model Indeterminacy and Unique Identification

With the factors and loadings unknown, the model is not uniquely identified. Therefore, Equation (2.1) can be expanded by an invertible matrix D, such that

$$y_t = \Lambda D D^{-1} f_t + e_t = \Lambda^* f_t^* + e_t \qquad \text{for all} \quad t \in \{1, \ldots, T\}, \tag{2.5}$$

[2] Y is assumed to be demeaned, otherwise an intercept parameter has to be included in the model, so f_t must be replaced by $[1' f_t']'$ and $\lambda_i \in \mathbb{R}^{K+1}$, where $\lambda_{i,1}$ is the intercept and $\lambda_{i,2:K}$ are the factor loadings.

and Λ^* and f_t^* for all $t \in \{1, \ldots, T\}$ are alternative choices for the loadings matrix and the factors. In fact, this indeterminacy allows for infinitely many alternative choices. This can also be seen from the likelihood function that obtains if the factors and errors are assumed to be normally distributed, i.e.

$$\begin{pmatrix} f_t \\ e_t \end{pmatrix} \sim f_N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & 0 \\ 0 & \Sigma \end{pmatrix} \right) \qquad \text{for all} \quad t \in \{1, \ldots, T\}. \tag{2.6}$$

The likelihood function is then

$$\mathcal{L}\left(\{y_t\}_{t=1}^T | \{f_t\}_{t=1}^T, \Lambda, \Sigma\right) = \prod_{t=1}^{T} (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (y_t - \Lambda f_t)' \Sigma^{-1} (y_t - \Lambda f_t) \right), \tag{2.7}$$

and also remains unchanged by the expansion of Λf_t.

The issue of indeterminacy can be solved in different ways, see e.g. Millsap (2001) for a discussion. For instance, constraining the top $K \times K$ section of Λ to the identity matrix I_K solves the model indeterminacy and does not require any constraints on the factors, which can therefore be arbitrarily scaled and correlated. Alternatively, the factors can be modeled as mutually orthogonal, which constrains the factor covariances to zero. In this case, the top $K \times K$ section of Λ is constrained to a lower triangular matrix with ones on the diagonal, which leaves room for the factors to be arbitrarily scaled. A third approach, which is discussed in the following, decomposes the indeterminacy issue into two problems, which are dealt with by separately constraining the factors and loadings. The first is the *scaling problem*, and the second is the *rotation problem*. The scaling problem can be solved by the additional assumption that $\Omega = I_K$, see e.g. Thurstone (1935) or Anderson and Rubin (1956), so the factors are uncorrelated and have unit scale. Integrating the factors out from Equation (2.7) then yields the marginalized likelihood

$$\mathcal{L}(\{y_t\}_{t=1}^T | \Lambda, \Sigma) = \prod_{t=1}^{T} |(\Sigma + \Lambda\Lambda')|^{-\frac{1}{2}} (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{1}{2} y_t' (\Sigma + \Lambda\Lambda')^{-1} y_t \right), \tag{2.8}$$

which, containing the expression $\Lambda\Lambda'$, remains subject to the rotation problem. As the matrix D in Equation 2.5 can in fact be any orthogonal matrix, which includes rotation matrices, but is not limited to them, the rotation problem would be more appropriately named *orthogonal transformation problem*. In the following, however, I will stick to the established terminology.

2.2.2 Solving the Rotation Problem

The rotation problem can be solved in several ways, some of which are discussed by Anderson and Rubin (1956). One approach is to constrain $\Lambda'\Lambda$ to a diagonal matrix with diagonal elements arranged in descending order,[3] another one is to constrain the elements of Λ above

[3] This identification scheme requires that the diagonal elements are different from each other, see Anderson and Rubin (1956).

the diagonal to zero, hence obtaining a lower triangular (LT) form for Λ.[4] Both sets of constraints, however, only guarantee a *local identification*, i.e. the values for Λ that satisfy them are not globally unique. If one Λ is known that satisfies any of the two constraints, others that also do can be found by flipping the sign of a subset of its columns.

A necessary and sufficient set of constraints for a *global identification* and thus a globally unique solution to the rotation problem is to fix all elements above the diagonal of Λ to zero, and all elements on the diagonal to positivity, which yields a positive lower triangular (PLT) loadings matrix. This can be motivated by Theorem A9.8 from Muirhead (1982), or using the QR decomposition, see e.g. Golub and van Loan (2013). In the following, the latter approach shall be taken. First, consider the formal definition of an orthogonal matrix:

Definition 2.2.1. *Orthogonal matrices*
A matrix $D \in \mathbb{R}^{K\times K}$*, where* $DD' = D'D = I_K$*, is an orthogonal matrix. Since* $\det(D) = \det(D')$ *and* $\det(I_K) = 1$*, it must hold that* $|\det(D)| = 1$*. If* D *is an orthogonal matrix, then* D' *also is.*

All orthogonal matrices together form the orthogonal group, which is defined as follows:

Definition 2.2.2. *The orthogonal and special orthogonal group*
All K*-dimensional orthogonal matrices are members of the orthogonal group* $O(K)$*. Those* K*-dimensional orthogonal matrices for which* $\det(D) = 1$ *are called special orthogonal matrices and are members of the special orthogonal group* $SO(K)$*, which is a subgroup of* $O(K)$*.*

The group law of $O(K)$, and also of $SO(K)$, is the matrix multiplication, hence the following holds:

Definition 2.2.3. *Products of orthogonal matrices*
The product of two orthogonal matrices is itself an orthogonal matrix, i.e. for $D_3 = D_1D_2$*, where* $D_1, D_2 \in O(K)$*,* $D_3 \in O(K)$*. This implies that every orthogonal matrix can also be written as the product of two other orthogonal matrices. Every special orthogonal matrix can be written as the product of two special orthogonal matrices or as the product of two orthogonal matrices with determinant -1. Every orthogonal matrix with determinant -1 can be written as a the product of two matrices, one of which has determinant +1 and one of which has determinant -1.*

Some other important properties of orthogonal matrices useful to understand the concept of orthogonal mixing are discussed in Section 2.4.

A QR decomposition of the transpose of Λ yields

$$\Lambda' = QR, \tag{2.9}$$

where R is an upper triangular matrix, and Q is an orthogonal matrix. The QR decomposition is not unique, as it is possible choose B, such that

$$\Lambda' = QBB^{-1}R, \tag{2.10}$$

[4] This identification scheme requires that the diagonal elements are different from zero, see Dunn (1973).

where B is a diagonal matrix whose nonzero entries can take values -1 or 1, which is called a reflection matrix.

Definition 2.2.4. *Reflection matrices*
Let $B = diag(b_1, \ldots, b_K)$ where $b_k \in \{-1, 1\}$ for $k \in \{1, \ldots, K\}$. Then B is a reflection matrix that reflects the k^{th} element of a $K \times 1$ vector about the k^{th} axis, thus reversing its sign, if and only if $b_k = -1$. Since $BB' = B'B = I_K$, reflection matrices are orthogonal matrices.

Note that by Definition 2.2.3, QB is also an orthogonal matrix. Unless in the case where at least one diagonal element of R is equal to zero, which implies a rank deficit for the top $K \times K$ submatrix of Λ, it is possible to choose B in such a way that $B^{-1}R = B'R = BR$ has strictly positive elements on the diagonal, which is achieved by setting $b_k = \text{sgn}(r_{k,k})$ for all $k \in \{1, \ldots, K\}$, where $r_{k,k}$ denotes the k^{th} diagonal elements of R. This provides the unique decomposition

$$\Lambda' = D\Lambda'_{PLT}, \tag{2.11}$$

with $D = QB$ orthogonal and Λ_{PLT} positive lower triangular. Therefore, the likelihood function in Equation (2.7) and the marginalized likelihood function in Equation (2.8) each have only a single global maximum if Λ is PLT. Constraining Λ accordingly is therefore sufficient to solve the rotation problem. All the maxima from the unconstrained likelihood can accordingly be mapped onto this single maximum using the orthogonal transformation in Equation (2.11).

2.3 Model Identification in Gibbs Sampling

Section 2.2 discussed the rotation problem in the context of maximum likelihood factor analysis and showed that it can be solved by maximizing a constrained instead of an unconstrained likelihood. In Bayesian factor analysis, estimates for the parameters and factors are obtained from their posterior distributions. The Bayes estimator as a function of the posterior distribution is then the estimator that minimizes the Bayes risk, or the Bayesian expected loss. Consequently, it depends on the selected loss function, see Berger (1985). It is commonplace to choose a quadratic loss function, which implies that the posterior mean serves as the Bayes estimator.

If the posterior distribution is not analytically tractable, it may still be possible to generate a sample from it using MC methods. The posterior moments of interest can then be estimated from these samples, see e.g. Kloek and van Dijk (1978). Bayesian factor analysis typically uses the MCMC method of Gibbs sampling, proposed by Geman and Geman (1984). The Gibbs sampler iteratively simulates the parameters of interest from their full conditional distributions, which generates a sample from the posterior distribution, see e.g. Casella and George (1992).

2.3.1 A Gibbs Sampler for the Static Factor Model

The Gibbs sampler for the static factor model discussed in this chapter largely follows the setup of Otrok and Whiteman (1998) and Kose et al. (2003), but omits the parameters governing the dynamics in the factors and the according filtering, or quasi-differencing, steps that are part of the dynamic single- and multi-factor models discussed there. The chosen prior distributions are conjugate and independent, where the prior distribution of the loadings is a matrix normal distribution that is the product of independent K-variate normal distributions, and the prior distribution of the idiosyncratic error covariances is an inverse Wishart distribution that is the product of independent univariate inverse gamma distributions, hence

$$\pi(\Lambda, \Sigma) = \pi(\Lambda)\pi(\Sigma) = \prod_{i=1}^{N} f_N(\lambda_i|\mu_{\lambda_i}, \Sigma_{\lambda_i}) \prod_{i=1}^{N} f_{IG}(\sigma_i^2|\underline{\alpha}_i, \underline{\beta}_i). \tag{2.12}$$

In order to enable the Gibbs sampler to generate a sample from the posterior density of the model parameters of interest Λ and Σ, the latent factors are likewise sampled and are used in a data augmentation step, see Tanner and Wong (1987). The prior distribution of the factors is a K-variate normal distribution,

$$\pi\left(\{f_t\}_{t=1}^T\right) = \prod_{t=1}^{T} f_N(f_t|\mu_f, \Sigma_f). \tag{2.13}$$

Now the Gibbs sampler for the static factor model proceeds as follows for every iteration $z \in \{1, \ldots, Z\}$:

1. Sample $\Lambda^{(z)}$ from its full conditional distribution $\Lambda^{(z)}|\{f_t^{(z-1)}\}_{t=1}^T, \Sigma^{(z-1)}; Y$.
2. Sample $\Sigma^{(z)}$ from its full conditional distribution $\Sigma^{(z)}|\{f_t^{(z-1)}\}_{t=1}^T, \Lambda^{(z)}; Y$.
3. Sample the factors from their full conditional distribution $\{f_t^{(z)}\}_{t=1}^T|\Lambda^{(z)}, \Sigma^{(z)}; Y$.

In the following, the superscript z denoting the iteration is omitted for simplicity. The full conditional distribution of the loadings is

$$g(\Lambda|\{f_t\}_{t=1}^T, \Sigma, \{y_t\}_{t=1}^T) = \prod_{i=1}^{N} (2\pi)^{-\frac{K}{2}} |\Sigma_{\lambda_i}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\lambda_i - \mu_{l_i})'\Sigma_{l_i}(\lambda_i - \mu_{l_i})\right), \tag{2.14}$$

where $\Sigma_{l_i} = \left(\sigma_i^{-2}\sum_{t=1}^T f_t f_t' + \Sigma_{\lambda_i}^{-1}\right)^{-1}$ and $\mu_{l_i} = \Sigma_{l_i}\left(\Sigma_{\lambda_i}^{-1}\mu_{\lambda_i} + \sigma_i^{-2}\sum_{t=1}^T f_t y_{it}\right)$, the full conditional distribution of the idiosyncratic variances is

$$g\left(\Sigma|\{f_t\}_{t=1}^T, \Lambda, \{y_t\}_{t=1}^T\right) = \prod_{i=1}^{N} \frac{b_i^{a_i}}{\Gamma(a_i)} (\sigma_i)^{-2a_i-1} \exp\left(-\frac{b_i}{\sigma_i^2}\right), \tag{2.15}$$

where $a_i = \frac{T}{2} + \underline{\alpha}_i$ and $b_i = \frac{1}{2}\sum_{t=1}^{T}(y_t - \lambda_i' f_t)^2 + \underline{\beta}_i$, and the full conditional distribution of the factors is

$$g\left(\{f_t\}_{t=1}^T|\Lambda,\Sigma,\{y_t\}_{t=1}^T\right) = \prod_{t=1}^{T}(2\pi)^{-\frac{K}{2}}|\Sigma_{f_t}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(f_t-\mu_{f_t})'\Sigma_{f_t}^{-1}(f_t-\mu_{f_t})\right), \quad (2.16)$$

where $\Sigma_{f_t} = (\Lambda'\Sigma^{-1}\Lambda + \Sigma_f^{-1})^{-1} = (\Lambda'\Sigma^{-1}\Lambda + I_K)^{-1}$ and $\mu_{f_t} = \Sigma_{f_t}(\Sigma_f^{-1}\mu_f + \Lambda'\Sigma^{-1}y_t) = \Sigma_{f_t}(\Lambda'\Sigma^{-1}y_t)$.[5]

2.3.2 Dealing with Model Indeterminacies in Bayesian Factor Analysis

In Bayesian factor analysis, the indeterminacies discussed in Section 2.2 are dealt with by choosing appropriate prior hyperparameters. The constraint imposed on the factors in ML factor analysis to settle the scaling issue accordingly enters the model by setting μ_f to a $K \times 1$ vector of zeros, and setting Σ_f to the K-dimensional identity matrix I_K.[6] This choice of hyperparameters does not guarantee that the posterior estimates of the factors are perfectly uncorrelated and have unit variance each.[7]

As for the rotation issue, the identification constraints known from ML factor analysis can be imposed by choosing fully informative prior distributions. Hence, it is possible to fix the top $K \times K$ section of Λ to the identity matrix, and sample the remaining model parameters conditional on this specification. This set of identification constraints is referred to as "DFM2" by Bai and Wang (2012) as they consider different identification schemes for dynamic factor models. It can be embedded in the Gibbs sampler setup above by using Dirac delta distributions $\delta_0(\lambda_{i,j})$ as priors for $i \neq j$ and $i \leq K$, and for $\delta_0(\lambda_{i,i} - 1)$ as priors for $i \leq K$.

For the normal priors used above, this can also be understood as the limiting case for $\mu_{\lambda_i} = u_i$, where u_i denotes the i^{th} canonical unit vector, and $\Sigma_{\lambda_i} \to 0_{K\times K}$ for $i \leq K$. This setup does not imply uncorrelated factors, however, so the prior hyperparameters Σ_f can also be non-diagonal. The corresponding identification constraint for uncorrelated factors, where Λ is lower triangular with ones on the diagonal is implemented accordingly, except that the Dirac delta distributions $\delta_0(\lambda_{i,j})$ are chosen as priors for $i < j$ and $i \leq K$ only. For the normal priors used above, this can be understood as a limiting case for $\mu_{\lambda_i,[i:K]} = u_1$ and Σ_{λ_i} for

[5] In the dynamic factor model, persistence parameters for the factors have to be added and the factors can either be drawn directly after quasi-differencing the parameters of interest, as implemented for a single-factor model by Otrok and Whiteman (1998) and for a multi-factor model by Kose et al. (2003), or by using the multi-move Gibbs sampler of Carter and Kohn (1994), as implemented by Bai and Wang (2012) and discussed in detail in Chapter 3, or by drawing the factors in a single sweep, using the approach proposed by Chan and Jeliazkov (2009) and used e.g. in Kaufmann and Schumacher (2013).

[6] As the factors are centered about zero, other appropriate choices for μ_f do not exist, whereas choosing a multiple of I_K for Σ_f is possible, albeit not in line with the identifying assumptions discussed in Section 2.2.

[7] The correlation between the factors, however, is very small in absolute value, and the scale of the factors depends on the hyperparameters chosen for Λ and the relation between the number of variables N and the number of observations T per variable. In some applications, e.g. in the supplementary program code for Bernanke et al. (2005) and for Koop and Korobilis (2010), the Gibbs sweeps of the factors are therefore postprocessed, which usually involves a demeaning, and sometimes a rescaling to unit variance before the sampler proceeds.

$i \leq K$, $\Sigma_{\lambda_i,[i:K,i:K]} \to 0_{(K+1-i)\times(K+1-i)}$, where $\mu_{\lambda_i,[i:K]}$ denotes the last $K+1-i$ elements of μ_λ and $\Sigma_{\lambda_i,[i:K,i:K]}$ denotes the $(K+1-i)\times(K+1-i)$ lower right section of Σ_{λ_i}. This identification constraint is used in the Bayesian analysis of Aguilar and West (2000).

Eventually, if the factors are uncorrelated and a scaling assumption has been introduced by choosing the prior hyperparameters μ_f and Σ_f appropriately, only the rotation problem remains to be solved. The result from ML factor analysis that constrains Λ to the set of positive lower triangular matrices to obtain a likelihood with a unique maximum has been introduced to Bayesian factor analysis using the Gibbs sampler by Geweke and Zhou (1996) and has been used in numerous applications since then. Bai and Wang (2012) refer to this identification scheme as "DFM1". It is implemented by constraining the elements of Λ above the diagonal to zero, or, equivalently, using Dirac delta distributions $\delta_0(\lambda_{i,j})$ as priors for $i < j$. For the elements on and below the diagonal with $i \leq K$, a normal prior with hyperparameters $\mu_{\lambda_i,[1:i]}$ and $\Sigma_{\lambda_i,[1:i,1:i]}$ is used, which is truncated below at zero along the i^{th} dimension, i.e. the drawn $\lambda_{i,i}$ are guaranteed to be strictly positive to meet the identification constraint. This identification scheme is analyzed for a static factor model in Section 2.9 and in an extensive simulation study for static and dynamic factor models in Chapter 3.

2.3.3 The Rotation Problem and Label Switching

An argument against using the PLT constraint resembles the major argument against ex-ante identification in Markov switching and mixture models, see e.g. Stephens (2000), Frühwirth-Schnatter (2001), Jasra et al. (2005) and Frühwirth-Schnatter (2006). Consider e.g. a univariate normal mixture model with two components, where the data $X = \{x_1, \ldots, x_n\}$ is generated as an i.i.d. sample from the density

$$f(x;\theta) = \pi_1 f_N(x|\mu_1, \sigma_1^2) + \pi_2 f_N(x|\mu_2, \sigma_2^2) \tag{2.17}$$

with parameter vector $\theta = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)'$, where $\pi_2 = 1 - \pi_1$, so the data likelihood is

$$\mathcal{L}(\theta; X) = \prod_{i=1}^{n} f(x_i;\theta). \tag{2.18}$$

The auxiliary variables $\Xi = \{\xi_1, \ldots, \xi_n\}$ with $\xi_i \in \{1,\ 2\}$ for all $i \in \{1, \ldots, n\}$ indicate which of the two components the observations are assigned to, and are therefore also denoted as labels, see Stephens (2000). This model is not uniquely identified: A permutation of the parameter vector $\nu(\theta)$, which exchanges the positions of π_1 and π_2, of μ_1 and μ_2, and of σ_1^2 and σ_2^2, accompanied by an exchange of the labels for every ξ_i for $i \in \{1, \ldots, n\}$ yields the same likelihood value. This has been pointed out by Redner and Walker (1984) and bears several similarities to the model identification issue in factor analysis. The labeling problem is resolved for an ML analysis e.g. by imposing a constraint like $\mu_1 < \mu_2$, $\sigma_1 > \sigma_2$ or a similar one.

If Gibbs sampling is used for inference to simulate from the posterior density of the parameters of interest, label switching may occur as the sampler proceeds. As a result, the generated Gibbs sequences for the parameter vector are not fit for inference, as e.g. the Gibbs sequence for μ_1 of length Z may in fact consist of Z_1 draws from μ_1, followed by Z_2 draws from μ_2 after a label switching, and again followed by $Z_3 = Z - Z_1 - Z_2$ draws from μ_1 after a second label switching, so only the first Z_1 and the last Z_3 elements of the full sample are in fact draws from the posterior distribution of μ_1. Therefore, identifiability constraints can be imposed to suppress the label switching. These constraints, which - being the equivalent to constraints in ML inference - hold in the global maximum of the posterior distribution, are fit to "break the symmetry of prior (and thus the posterior)", see Stephens (2000). The constraints, however, do not hold everywhere, so they become "ineffective in removing the symmetry in the posterior distribution."

This can be illustrated by a small example. Consider two bivariate parameters of interest θ_1 and θ_2, e.g. the mean vectors of a two-component bivariate mixture distribution, with pdfs

$$\theta_1 \sim f_N\left(\begin{pmatrix}3\\1\end{pmatrix}, \begin{pmatrix}2 & 0.4\\0.4 & 1\end{pmatrix}\right) \quad \text{and} \quad \theta_2 \sim f_N\left(\begin{pmatrix}4\\-2\end{pmatrix}, \begin{pmatrix}2 & -0.8\\-0.8 & 0.5\end{pmatrix}\right). \tag{2.19}$$

Let the distributions of θ_1 and θ_2 be unknown, but assume that it is possible to generate a sample from them, where the labels of θ_1 and θ_2 may switch in the process of sampling. Figure 2.1 shows a sample from the unknown distribution of θ_1 in the top panel of the first column, and a sample from the unknown distribution of θ_2 in the bottom panel of the first column. Next, consider the available samples, generated subject to label switching. The purported sample for θ_1 is shown in the top panel of the second column, and the resulting purported sample for θ_2 is shown in the bottom panel of the second column. Now there are different ways to prevent the label switching in the process of sampling: If $\theta_{1,1} < \theta_{2,1}$ is assumed to hold, in a pair of draws from both distributions, the one with the smaller first element must originate from the distribution of θ_1, whereas the one with the larger first element must originate from the distribution of θ_2. If this constraint is imposed, the purported sample from θ_1 looks as shown in the top panel of the third column, and the purported sample from θ_2 looks as shown in the bottom panel of the third column. The plots show that this constraint, where the discrimination hinges on the first element of θ_1 and θ_2, does not work well to prevent the label switching and can thus be considered *ineffective*. Conversely, if $\theta_{1,2} > \theta_{2,2}$ is assumed to hold, in a pair of draws from both distributions, the one with the larger second element must originate from the distribution of θ_1, whereas the one with the smaller second element must originate from the distribution of θ_2. If this constraint is imposed, the purported sample from θ_1 looks as shown in the top panel of the fourth column, and the purported sample from θ_2 looks as shown in the bottom panel of the fourth column. The plots show that this alternative constraint, where the discrimination hinges on the second element of θ_1 and θ_2, works well to prevent the label switching and can thus be considered *effective*. There may, however, not always exist constraints that succeed in preventing the label switching, and even if they exist, it is generally not easy to find them, particularly in models with many parameters.

2.3.4 The Ordering Problem in Bayesian Factor Analysis

The constraints imposed to prevent label switching are similar to the PLT constraint imposed to overcome the rotation indeterminacy. The major similarity can be motivated as follows: When applying the Gibbs sampler to simulate from the posterior distribution of Λ in a factor model, the zero constraints that are supposed to prevent the factors from switching their labels may not suffice. Consider a model with $K = 2$ factors, with model parameters

$$\Lambda = \begin{pmatrix} \lambda_{\cdot,1} & \lambda_{\cdot,2} \end{pmatrix}, \quad \Sigma \quad \text{and} \quad F = \begin{pmatrix} f_{\cdot,1} & f_{\cdot,2} \end{pmatrix}, \tag{2.20}$$

where $\lambda_{\cdot,k}$ denotes the k^{th} column vector of Λ, containing the loadings on the k^{th} factor, and $f_{\cdot,k}$ denotes the k^{th} column vector of F, containing the k^{th} factor itself.

To guarantee that the posterior distribution has only a single global maximum, the PLT constraint requires that $\lambda_{1,2} = 0$, $\lambda_{1,1} > 0$ and $\lambda_{2,2} > 0$.[8] Now assume that the factors $f_{\cdot,1}$ and $f_{\cdot,2}$ switch their labels, and the column vectors of Λ, $\lambda_{\cdot,1}$ and $\lambda_{\cdot,2}$, likewise switch their labels. This can be expressed by choosing for the matrix D in Equation 2.5 a permutation matrix, which exchanges the elements of vectors they are applied to. Assigning each element k to a new position $\nu(k)$, it can also be understood as a relabeling of the respective elements:

Definition 2.3.1. *Permutation matrices*
Let $P \in \mathbb{R}^{K \times K}$, *where* $P = \begin{pmatrix} u_{\nu(1)}, \ldots, u_{\nu(K)} \end{pmatrix}'$, *where* u_k *denotes the* k^{th} K*-dimensional canonical unit vector and* $\{\nu(1), \ldots, \nu(K)\}$ *denotes a permutation of integers* $\{1, \ldots, K\}$. *Then* P *is a permutation matrix that moves the* k^{th} *element of a* $K \times 1$ *vector into the* $\nu(k)^{th}$ *position. Every permutation matrix is an orthogonal matrix. Denote a permutation matrix with* $\det(P) = 1$ *as an even permutation matrix and a permutation matrix with* $\det(P) = -1$ *as an odd permutation matrix.*

Relabeling the factors and corresponding column vectors of Λ yields an alternative posterior distribution. The PLT constraint has the effect that the maximum of this distribution under label switching is excluded from the sample space. To be effective, however, the constraint should rule out sampling from the distribution under label switching altogether. At the same time, it should not prevent sampling from the desired distribution under the original labeling. In other words, it should redistribute as much of the probability mass as possible that it is supposed to redistribute, and as little of the probability mass as possible that it is not supposed to redistribute. To best achieve this, the unknown posterior distribution of Λ should have as little mass as possible in points with $\lambda_{1,1} \leq 0$ and $\lambda_{2,2} \leq 0$ when $\lambda_{1,2}$ is constrained to zero. Such information to base a decision whether or not the PLT constraint is effective on is not available ex ante. Now consider an $N \times N$ permutation matrix O that is premultiplied to the factor model representation from Equation 2.1. This yields

$$Oy_t = O\Lambda f_t + Oe_t \qquad \text{for all} \quad t \in \{1, \ldots, T\}, \tag{2.21}$$

[8] To prevent label switching of the factors and corresponding column vectors of Λ, the LT constraint, omitting the positivity restrictions on $\lambda_{i,i}$ for $i \in \{1, \ldots, K\}$ suffices.

where the reordered data $YO = (Oy_1, \ldots, Oy_T)'$ have the same factors F as in the original model, albeit with the new parameters $\Lambda^* = O\Lambda$ and $\Sigma^* = O\Sigma O$. It can be seen that constraining the sample space of $\lambda_{1,1}$, $\lambda_{1,2}$ and $\lambda_{2,2}$ may lead to a quite different behavior of the Gibbs sampler than constraining the sample space of $\lambda^*_{1,1}$, $\lambda^*_{1,2}$ and $\lambda^*_{2,2}$, and hence produce quite different inference results. This phenomenon occurring whenever the PLT constraint is chosen for model identification has been named the *ordering problem* observed and discussed e.g. by Lopes and West (2004), Carvalho et al. (2008) and Frühwirth-Schnatter and Lopes (2012). The ordering problem is illustrated by means of an example in Section 2.9 and analyzed in more detail in Chapter 3. Referring to the particular importance of the variables in the first K positions, Carvalho et al. (2008) calls these variables the *founders* of the factors. This seems to be a more accurate description of the issue of choosing an effective PLT identification constraint, as once the variables in the first K positions have been chosen, the ordering of the remaining variables does not affect the behavior of the Gibbs sampler anymore.[9]

2.3.5 Unconstrained Gibbs Sampling

Due to the difficulties arising when the constraints intended to prevent label switching in Markov switching and mixture models are ineffectively chosen, Stephens (2000) proposes to omit these constraints and allow for label switching during the Gibbs sampling process. In a subsequent step, all elements of the sample undergo an iterative relabeling procedure, which is described in more detail in Section 2.5. In a similar fashion, the issues arising if the factor founders are not effectively chosen and the difficulty of appropriately choosing them while the properties of the posterior distribution that are vital to choose them well are still unknown can be circumvented by not imposing any constraints to solve the rotation problem ex ante. This leads to orthogonal mixing during the sampling process, which is described in detail in Section 2.4. The algorithm used in the postprocessing step corresponding to the relabeling in Markov switching and mixture models is described in Section 2.6.

An unconstrained Gibbs sampler resembles the Gibbs sampler described in Subsection 2.3.1, but it is run without any of the constraints from Subsection 2.3.2 intended to solve the rotation problem ex ante. As a result, the factor space is orthogonally transformed during the process of Gibbs sampling. The Gibbs sequences obtained for Λ and F are therefore obtained not from the posterior distribution of interest, but from an orthogonal mixture of this distribution. The unconstrained Gibbs sampler is therefore called *orthogonally mixing*.

2.4 Orthogonal Mixtures

In Bayesian factor analysis using the Gibbs sampler as discussed in Section 2.3, the unconstrained sampler described in Subsection 2.3.5 generates a Gibbs sequence for Σ that

[9] Note that reordering the variables and imposing the PLT constraint afterwards is equivalent to leaving the ordering unchanged and imposing the zero and positivity constraints on different elements of Λ.

is a sample from the posterior distribution of Σ and that can therefore directly be used for inference. This is due to the fact that Σ is unaffected by the rotation problem, since Λ and F enter its full conditional distribution, given in Equation (2.15), only as a product. The Gibbs sequences generated for Λ and F, however, are samples from *orthogonal mixtures* of the posterior distribution, which evolve due to orthogonal transformations of the sample space for Λ that occur during the sampling process.[10] These orthogonal transformations have to be reversed in a subsequent postprocessing step. The orthogonal transformations include label switching of some or all factors and the corresponding factor loadings, as discussed above, a sign switching of one or more factors and the corresponding loadings, and rotations of the sample space of Λ. To understand better what the unconstrained sampler does, it is useful to look at the properties of orthogonal matrices, see Definition 2.2.1, and of the orthogonal group, see Definition 2.2.2. Details on the matrix algebra in this section can be found e.g. in Artin (1991), the matrix decompositions discussed here are explained in more detail e.g. in Bernstein (2009) and Golub and van Loan (2013).

2.4.1 Properties of Orthogonal Matrices

The orthogonal group of matrices of dimension $K \times K$, denoted as $O(K)$, is a subgroup of the Euclidean group $E(K)$. Orthogonal matrices inherit the Euclidean group's isometry property, i.e. the property of preserving distances, while lacking the translational property, i.e. the capability of describing rigid motions. The isometry property implies that lengths of vectors and angles between them are preserved under a joint transformation of the vectors by an orthogonal matrix, hence:

Definition 2.4.1. *Length- and angle preserving property*
Orthogonal matrices are length- and angle-preserving. Hence, if $v_1, v_2 \in \mathbb{R}^K$ and $D \in O(K)$, then $\|Dv_i\| = \|v_i\|$ for $i \in \{1, 2\}$ and the angle between v_1 and v_2 is identical to the angle between Dv_1 and Dv_2. This can be seen from the cosine formula

$$\cos(\gamma) = \frac{(Dv_1)'(Dv_2)}{\|Dv_1\|\|Dv_2\|} = \frac{v_1' D' D v_2}{\|v_1\|\|v_2\|} = \frac{v_1' v_2}{\|v_1\|\|v_2\|}, \tag{2.22}$$

where $\|v_1\|$ denotes the Euclidean norm of the vector v_1.

While for D with $\det(D) = -1$, D produces a mirror image of the vectors it is applied to, D with $\det(D) = 1$ does not perform such a reflection and thus preserves their orientation:

Definition 2.4.2. *Orientation preserving property*
Special orthogonal matrices are also orientation-preserving.

Preserving the orientation of the vectors it is applied to, having the isometry property and lacking the translational property, all a special orthogonal matrix can do is to perform a rotation about a fixed origin. Hence, a special orthogonal matrix can also be called that way:

[10] F is an augmented parameter, whose sample space depends directly on that of Λ. All transformations of the sample space of Λ therefore also directly affect the sample space of F.

Definition 2.4.3. *Rotation matrices*
Special orthogonal matrices can also be called rotation matrices. For $K = 2$, a rotation matrix around an angle γ is defined as

$$Q(\gamma) = \begin{pmatrix} \cos(\gamma) & -\sin(\gamma) \\ \sin(\gamma) & \cos(\gamma) \end{pmatrix}. \tag{2.23}$$

The product of two rotation matrices $Q_1 = Q(\gamma_1), Q_2 = Q(\gamma_2)$, where $Q_1, Q_2 \in \mathbb{R}^{K\times K}$ is likewise a rotation matrix, where for $K = 2$, $Q_3 = Q_1 Q_2 = Q(\gamma_1 + \gamma_2)$.

Rotation matrices are of crucial importance for the rotation problem and thus the most obvious type of orthogonal matrix D in Equation (2.5). Unlike a PLT constrained Gibbs sampler, the unconstrained Gibbs sampler does not suppress rotations, so inbetween every two subsequent Gibbs iterations z and $z + 1$, a rotation around an angle γ, where γ is small in magnitude, takes place. The postprocessing procedure proposed in Section 2.6 provides estimates for D, from which estimates for γ can be derived using Theorem 2.4.1. This is illustrated in the examples given in Section 2.9.

It is also possible to rotate only about a subset of the K axes with at least two elements, e.g. about 2 out of the K axes for $K \geq 2$. Hence there exist $\binom{K}{2}$ such pairs of axes:

Definition 2.4.4. *Givens rotation matrices*
A rotation matrix $G \in \mathbb{R}^{K\times K}$ for $K > 2$ is called a Givens rotation matrix if it performs a rotation only about a single pair of axes k_1 and k_2. Hence

$$g_{i,j} = \begin{cases} -\sin(\gamma) & \text{if} \quad i = k_1 \ \wedge \ j = k_2 \\ \sin(\gamma) & \text{if} \quad i = k_2 \ \wedge \ j = k_1 \\ \cos(\gamma) & \text{if} \quad (i = k_1 \ \wedge \ j = k_1) \ \vee \ (i = k_2 \ \wedge \ j = k_2) \\ 1 & \text{if} \quad i = j \ \wedge \ i \neq k_1 \ \wedge \ i \neq k_2 \\ 0 & \text{otherwise.} \end{cases} \tag{2.24}$$

If all possible pairs of axes are considered, there exists a representation of any rotation matrix in terms of Givens rotation matrices. This is a special case of the decomposition in Theorem 2.4.1:

Definition 2.4.5. *Givens decomposition*
Every rotation matrix Q can be written as the product of at most $K(K-1)/2$ Givens rotation matrices $\{G_h\}_{h=1}^{K(K-1)/2}$.

From Definitions 2.4.3 and 2.2.2, it follows that a matrix with determinant 1 is a rotation matrix, so this holds also for a subgroup of the permutation matrices from Definition 2.3.1:

Definition 2.4.6. *Even permutation matrices*
Every even permutation matrix is also a rotation matrix.

Consider for instance the permutation $\{\nu(1), \nu(2), \nu(3)\} = \{2, 3, 1\}$, which according to Definition 2.3.1 yields the matrix $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$. This matrix has determinant 1 and is therefore a rotation matrix. The same matrix can be obtained by a rotation matrix that by Definition 2.4.4 is decomposable into pairwise axis rotations. A possible set of angles for these rotations, which can be found by applying Theorem 2.4.1, is $\{\gamma_1, \gamma_2, \gamma_3\} = \left\{\pi, \frac{\pi}{2}, \frac{\pi}{2}\right\}$.

In order to change the sign of an element of a vector, a reflection matrix as in Definition 2.2.4 is required. Note that a reflection about two axes k_1 and k_2 is the same as a Givens rotation about the same pair of axes with $\gamma = \pi$. Consequently, it follows from Definition 2.4.5 that:

Definition 2.4.7. *Even reflection matrices*
In a reflection matrix $R = diag(r_1, \ldots, r_K)$, define ρ as the set of indices k for which $r_k = -1$. Denote R as an even reflection matrix if $\frac{|\rho|}{2} \in \mathbb{Z}$ and hence $\det(R) = 1$, otherwise denote R as an odd reflection matrix. Every even reflection matrix is also a rotation matrix.

One particular orthogonal matrix is the identity matrix, which falls into many of the subcategories of orthogonal matrices discussed here:[11]

Definition 2.4.8. *Identity matrix*
The identity matrix is an orthogonal matrix. Since $\det(I_K) = 1$, it is a special orthogonal matrix. For $K = 2$, it is also a rotation matrix with $\gamma = 0$, for $K > 2$, it is a the product over a single Givens rotation matrix with $\gamma_1 = 0$. Moreover, it is a permutation matrix with $\nu(k) = k$ for all $k \in \{1, \ldots, K\}$ and a reflection matrix with $\rho = \emptyset$.

The following theorem is related to a decomposition of orthogonal matrices given in Anderson et al. (1987), but uses a slightly different parametrization. In particular, the angles γ are defined over a range from $-\pi$ to π, instead of $-\pi/2$ to $\pi/2$, which reduces the number of reflection parameters from K to 1, hence this representation is more parsimonious and, having almost exclusively parameters that live in the continuous space, is more easy to handle in optimizations. Conversely, the extended range of the angles γ results in multiple representations of the same orthogonal matrix.

Theorem 2.4.1. *Decomposition of any orthogonal matrix*
Every orthogonal matrix $D \in \mathbb{R}^{K \times K}$ can be decomposed into $\binom{K}{2}$ Givens rotation matrices and one axis reflection matrix.

The proof of Theorem 2.4.1 can be found in Appendix 2.A.

2.4.2 Orthogonal Mixture Distributions

Having discussed some properties of orthogonal matrices in Subsection 2.4.1, a definition of orthogonal mixture distributions can now be given.

[11] The identity matrix is generally a trivial case for D barely worth mentioning. In the proof for Theorem 2.4.1, however, its property as a reflection matrix with $\rho = \emptyset$ is referred to.

Definition 2.4.9. *Finite orthogonal mixture distribution*
A finite orthogonal mixture distribution is a mixture distribution with probability density function $\sum_{i=1}^{m} \pi_i f(x;\theta_i)$*, where all sets of parameters for the individual mixture components* $\theta_i = \theta(D_i)$ *are functions of orthogonal matrices* D_i*, which have the effect that the* x *are accordingly orthogonally transformed into* xD_i*. Due to Definition 2.2.3, the pdf of a finite orthogonal mixture distribution is*

$$h(x;\{\theta_i\}_{i=1}^m) = h(x;\theta,\{D_i\}_{i=1}^m,\{\pi_i\}_{i=1}^{m-1}), \tag{2.25}$$

i.e. the pdf can be parameterized in terms of a single set of parameters θ*,* m *orthogonal matrices* $\{D_i\}_{i=1}^m$ *and* $m-1$ *mixture proportions* $\{\pi_i\}_{i=1}^{m-1}$*. The parameters* θ *are the parameters of the underlying orthogonally invariant distribution, see Definition 2.4.11. By Definition 2.2.3, it is possible to choose an orthogonal matrix* D *and transform* θ *into* $\theta(D)$ *and all matrices* D_i *into* $D'D_i$ *for* $i \in \{1,\dots,m\}$*. This yields a representation of the same finite orthogonal mixture distribution.*

Consider for instance the example of label switching discussed in Subsection 2.3.3. In the case of a mixture with two components, a Gibbs sampler that allows for label switching therefore produces a sample from an orthogonal mixture distribution with $m = 2$, with orthogonal matrices $D_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $D_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. This example is discussed in more depth in Section 2.5.

Definition 2.4.10. *Infinite orthogonal mixture distribution*
An orthogonal mixture distribution with pdf

$$h(x;\theta,D) \tag{2.26}$$

where the orthogonal matrix D *is a random variable, whose distribution has the pdf*

$$g(D;\psi), \tag{2.27}$$

is an infinite orthogonal mixture distribution. $h(x;\theta,\psi)$ *is the pdf of a compound distribution with*

$$h(x,\theta,\psi) = \int_{D\in\mathcal{D}} f(x;\theta,D)g(D;\psi)dD, \tag{2.28}$$

where $\mathcal{D}$ *denotes the domain of* D *and* ψ *denotes a vector of parameters that govern the distribution of* D*.*

The unconstrained Gibbs sampler for the static factor model generates a sample from an infinite mixture distribution, i.e. the number of orthogonal matrices D is infinitely large. While limiting the choice of D to permutation matrices, as in the case of label switching, would restrict the maximum number of mixture components to $K!$, allowing rotation matrices for D results in a different $D^{(z)}$ for every Gibbs sweep $z \in \{1,\dots,Z\}$. If the distributional law of the D is not known, the compound distribution in Equation (2.28) is intractable. As

a remedy in this case, the $D^{(z)}$ can be considered as the parameters D_i of a finite orthogonal mixture distribution with mixture proportions $\pi_i = \frac{1}{Z}$ for every $i \in \{1, \ldots, m\}$ and $m = Z$, and can be estimated for every $z \in \{1, \ldots, Z\}$, using the algorithm proposed in Section 2.6.

Definition 2.4.11. *Orthogonally invariant distribution*
If the D_i and π_i of a finite orthogonal mixture distribution $h(x; \theta, \{D_i\}_{i=1}^m, \{\pi_i\}_{i=1}^{m-1})$ are unknown, but a valid representation of θ is known, $h(\tilde{x}; \theta, I_K, 1)$ is the underlying orthogonally invariant distribution with the single possible choice for D being the identity matrix. Accordingly, if the distributional law of the D for an infinite orthogonal mixture distribution is unknown, but a valid representation of θ is known, $f(\tilde{x}; \theta, I_K)$ is the underlying orthogonally invariant distribution with the entire probability mass of D being concentrated in I_K.

The underlying orthogonally invariant distribution in the sense of Definition 2.4.11 is not unique. If some orthogonal matrix D is chosen to transform θ in the finite orthogonal mixture distribution into $\theta(D)$, and the matrices D_i into $D'D_i$ for $i \in \{1, \ldots, m\}$, a different valid representation of the underlying orthogonal mixture distribution with $\theta(D)$ is obtained.

Definition 2.4.12. *Orthogonally mixed sample*
A random sample $\{x_s\}_{s=1}^S$ from a finite or infinite orthogonal mixture distribution as in Definitions 2.4.9 and 2.4.10 is an orthogonally mixed sample. Unless stated otherwise, no particular ordering is assumed for the S elements of the sample.

In Definition 2.4.9, $\{D_i\}_{i=1}^m$ and $\{\pi_i\}_{i=1}^{m-1}$ are considered as parameters of a mixture distribution, whereas Definition 2.4.10 allows for D to be a random variable itself. When considering orthogonally mixed samples that evolve as a Markov chain of length Z, D can be understood as a latent state variable, and $D^{(z)}$ as its realization in the z^{th} element of the Markov chain. The realization of the unobservable variable D, which is $D^{(z)}$, directly affects $\theta^{(z)} = \theta(D^{(z)})$, which in turn affects the realization of the observable variable x, which is $x^{(z)}$.

Definition 2.4.13. *Orthogonal mixing as a hidden Markov process*
An orthogonally mixed sample as in Definition 2.4.12 can be the outcome of a Markov process, where every realization of x in the sequence of observations is generated conditional on θ transformed by a specific latent D. D is a state variable and follows a hidden Markov process.

Whenever an orthogonally mixed sample is at hand, the task is therefore twofold: First, a representation of the orthogonally invariant distribution described in Definition 2.4.11 shall be found, and second, the orthogonal matrix D_s that θ was transformed by to produce the sample element x_s shall be determined. In the case of the latent factor model, the orthogonally mixed sample is obtained as a Markov chain, so D is a latent state variable with realization $D^{(z)}$ in the z^{th} iteration of the Gibbs sampler, as described in Definition 2.4.13 and is to be estimated from the observable $\Lambda^{(z)}$. It will be seen in the following that these two tasks are interwoven in such a way that it is possible to obtain estimates for both of them in a joint procedure.

2.5 Orthogonal Mixing and Label Switching

The following examples are chosen to illustrate the concept of orthogonal mixture distributions from Definition 2.4.9. In each example, an orthogonally mixed sample as in Definition 2.4.12 is considered, where the properties of the underlying orthogonally invariant distribution as in Definition 2.4.11 and the properties of the orthogonal mixing process vary. The examples include well-known concepts such as label switching and mixture distributions.

2.5.1 Examples of Sign and Label Switching

The following examples are exclusively dealing with normal distributions for illustrative purposes, but can be generalized to many other elliptical distributions. Non-elliptical distributions are not discussed - their properties under orthogonal mixing are the same as described above, but they are more difficult to handle and therefore left for future investigation.

Example 2.5.1. *First consider a sample $\{x_s\}_{s=1}^S$ from a univariate normal distribution with $\theta = \{\mu; \sigma^2\}$ and $D_s \in \{-1, 1\}$. Hence the result of the orthogonal mixing is a two-component mixture with pdf*

$$h(x) = \pi f_N(-\mu, \sigma^2) + (1-\pi) f_N(\mu, \sigma^2), \tag{2.29}$$

which depends on the three parameters $\theta = \{\mu, \sigma^2, \pi\}$, where $\pi = P(D_s = -1)$. The parameters can be estimated as in a standard mixture model, see e.g. Frühwirth-Schnatter (2006).

Example 2.5.2. *Now assume that $\{x_s\}_{s=1}^S$ comes from an n-variate normal distribution with diagonal covariance matrix, hence $x_s = [x_{s,1}, \ldots, x_{s,n}]'$. The resulting two-component mixture then has the pdf*

$$h(x) = \pi f_N(-\mu, \Sigma) + (1-\pi) f_N(\mu, \Sigma). \tag{2.30}$$

This mixture distribution can be decomposed into n individual bivariate mixture distributions

$$h(x_j) = \pi f_N(-\mu_j, \sigma_j^2) + (1-\pi) f_N(\mu_j, \sigma_j^2) \quad \text{for } j \in \{1, \ldots, n\}, \tag{2.31}$$

so the parameter vector is $\theta = \{\{\mu_j\}_{j=1}^n, \{\sigma_j^2\}_{j=1}^n, \pi\}$. An increase in n generally makes the estimation of the D_s from the x_s easier and leads to more accurate estimates for θ.[12]

Example 2.5.3. *Next, consider a case where $\{x_s\}_{s=1}^S$ follows a bivariate normal distribution, hence $x_s = [x_{s,1}, x_{s,2}]'$. The mixing occurs in the same way as before, but individually for x_1 and x_2, so $D_s \in \left\{ I_2, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, -I_2 \right\} \subset O(2)$. Hence D_s is either the identity*

[12] An additional x_j does not provide any additional information if $\mu_j = 0$, however.

or flips the signs of one or both elements of x_s around and is therefore a reflection matrix as in Definition 2.2.4. The matrices I_2 and $-I_2$ are even reflection matrices and therefore also rotation matrices, see Definition 2.4.7, and the identity matrix I_2 is likewise a reflection matrix, see Definition 2.4.8. The parameter vector is therefore $\theta = \{\mu', vech(\Sigma)', \pi_1, \pi_2, \pi_3\}$, and the result of the orthogonal mixing is a four-component mixture distribution with pdf

$$h(x) = \pi_1 f_N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \Sigma\right) + \pi_2 f_N\left(\begin{pmatrix}-\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & -\sigma_{1,2}\\ -\sigma_{2,1} & \sigma_2^2\end{pmatrix}\right)$$
$$+ \pi_3 f_N\left(\begin{pmatrix}\mu_1\\ -\mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & -\sigma_{1,2}\\ -\sigma_{2,1} & \sigma_2^2\end{pmatrix}\right) + \left(1 - \sum_{i=1}^{3}\pi_i\right) f_N(-\mu, \Sigma). \tag{2.32}$$

If the sample is obtained as the outcome of a Markov process, the hidden state of the system in the sense of Definition 2.4.13 thus involves a sign switching that occurs whenever $D^{(z)} \neq D^{(z-1)}$. After estimating $D^{(z)}$ for all $x^{(z)}$, the mixture proportions π_1, π_2 and π_3, but also the 4×4 transition matrix of the Markov process can be estimated.

Example 2.5.4. *Now consider an example of label switching. Assume again that x_s follows a bivariate normal distribution, but let the mixing matrices $D_s \in \left\{I_2, \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\right\} \subset O(2)$, hence D_s is a permutation matrix, see Definition 2.3.1. The identity matrix is included in the set of permutation matrices for the $\mathbb{R}^2$, and it is also the only even permutation matrix, see Definition 2.4.6. The resulting mixture distribution in this case has pdf*

$$h(x) = \pi f_N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \Sigma\right) + (1-\pi) f_N\left(\begin{pmatrix}\mu_2\\ \mu_1\end{pmatrix}, \begin{pmatrix}\sigma_2^2 & \sigma_{2,1}\\ \sigma_{1,2} & \sigma_1^2\end{pmatrix}\right). \tag{2.33}$$

The hidden state of the system in the sense of Definition 2.4.13 thus involves a label switching that occurs whenever $D^{(z)} \neq D^{(z-1)}$. After estimating $D^{(z)}$ for all $x^{(z)}$, the mixture proportion π and the 2×2 transition matrix of the Markov process can be estimated.

Examples 2.5.3 and 2.5.4 can be generalized in a similar way as Example 2.5.1 is generalized in Example 2.5.2. Let $\{x_j\}_{j=1}^n$ follow independent K-variate normal distributions with mean vectors $\{\mu_j\}_{j=1}^n$ and covariance matrices $\{\Sigma_j\}_{j=1}^n$. Then $X = [x_1, \ldots, x_n]'$ is an $n \times K$ matrix with

$$\text{vec}(X') \sim N\left(\begin{pmatrix}\mu_1\\ \vdots\\ \mu_n\end{pmatrix}, \begin{pmatrix}\Sigma_1 & 0 & \ldots & 0\\ 0 & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \ldots & 0 & \Sigma_n\end{pmatrix}\right), \tag{2.34}$$

i.e. a conformable vectorization of X follows a multivariate normal distribution. Denote the $n \times K$ matrix of stacked transposed mean vectors $M = (\mu_1, \ldots, \mu_n)'$, and the $nK \times nK$ block-diagonal matrix of covariance matrices

$$\bar{\Sigma} = \begin{pmatrix} \Sigma_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \Sigma_n \end{pmatrix}. \tag{2.35}$$

Then Equation (2.34) can be written as

$$\operatorname{vec}(X') \sim N\left(\operatorname{vec}(M'), \bar{\Sigma}\right), \tag{2.36}$$

which resembles that of a matrix-normal distribution discussed e.g. in Dawid (1981), Dutilleul (1999), and Glanz and Carvalho (2013).[13]

In the following, denote the element of the sample corresponding to X_s from the underlying orthogonally invariant distribution as $\tilde{X}_s$. Then $X_s = \tilde{X}_s D_s$, where $\tilde{X}_s$ and D_s are latent, and X_s is observable. Let $\tilde{M} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_n)'$ and $\tilde{\bar{\Sigma}}$ denote the mean and covariance of the underlying orthogonally invariant distribution, where $\tilde{\bar{\Sigma}}$ is defined equivalent to $\bar{\Sigma}$ in Equation (2.35). The respective orthogonally transformed moments are then $M_s = \tilde{M} D_s = (D_s' \tilde{\mu}_1, \ldots, D_s' \tilde{\mu}_n)'$ and $\bar{\Sigma}_s$ with $\{\Sigma_{j,s}\}_{j=1}^n = \{D_s' \tilde{\Sigma}_j D_s\}_{j=1}^n$ arranged in block-diagonal form. If the D_s were known, it would therefore be possible to transform the X_s back into the $\tilde{X}_s$, which could then be used to estimate $\tilde{M}$ and $\tilde{\bar{\Sigma}}$.

The distribution of the X_s is an orthogonal mixture distribution. Conditional on the D_s, it can be written in terms of Equation (2.34) as

$$\begin{aligned} \operatorname{vec}(X_s' | D_s) &\sim N\left(\operatorname{vec}(D_s' \tilde{M}'), \begin{pmatrix} D_s' \Sigma_1 D_s & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & D_s' \Sigma_n D_s \end{pmatrix}\right) \\ &= N\left(\operatorname{vec}\left((I_n \otimes D_s')\tilde{M}\right), (I_n \otimes D_s')\tilde{\bar{\Sigma}}(I_n \otimes D_s)\right). \end{aligned} \tag{2.38}$$

Accordingly, the underlying orthogonally invariant distribution of the $\tilde{X}_s$ as per Definition 2.4.11 can be written as

$$\operatorname{vec}(\tilde{X}_s') \sim N\left(\operatorname{vec}(\tilde{M}'), \bar{\Sigma}\right). \tag{2.39}$$

[13] The matrix-normal distribution, however, has the form

$$\operatorname{vec}(X') \sim N\left(\operatorname{vec}(M'), \Sigma_r \otimes \Sigma_c\right), \tag{2.37}$$

with row covariance matrix $\Sigma_r \in \mathbb{R}^{n \times n}$ and column covariance matrix $\Sigma_c \in \mathbb{R}^{K \times K}$. Here, however, instead of the Kronecker product $\Sigma_r \otimes \Sigma_c$, the covariance matrix is a block diagonal matrix of the different column covariance matrices, while $\Sigma_r = I_n$, because of the mutual independence of the x_j for $j \in \{1, \ldots, n\}$.

Example 2.5.5. *Analogous to Example 2.5.4, consider the D_s in Equation (2.38) being exclusively permutation matrices, see Definition 2.3.1. This yields $K!$ possible choices for D_s and hence a finite orthogonal mixture with at most $K!$ components, see Definition 2.4.9.*

The label switching example discussed at the end of Section 2.3.3 with the two parameters of interest θ_1 and θ_2 corresponds to Example 2.5.5 with $K = 2$ and $n = 2$ with the underlying orthogonally invariant distribution

$$\begin{pmatrix} \theta_{1,1} \\ \theta_{1,2} \\ \theta_{2,1} \\ \theta_{2,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 3 \\ 1 \\ 4 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 & 0.4 & 0 & 0 \\ 0.4 & 1 & 0 & 0 \\ 0 & 0 & 2 & -0.8 \\ 0 & 0 & -0.8 & 0.5 \end{pmatrix} \right) \tag{2.40}$$

and $D_s \in \left\{ I_2, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}$.

Example 2.5.6. *Analogous to Example 2.5.3, consider the D_s in (2.38) being exclusively reflection matrices, see Definition 2.2.4. This yields 2^K possible choices for D_s and hence a finite orthogonal mixture with at most 2^K components, see Definition 2.4.9.*

Example 2.5.7. *Combining Examples 2.5.5 and 2.5.6, consider D_s in (2.38) being reflection or permutation matrices, or combinations thereof. This yields $2^K K!$ possible choices for D_s and hence a finite orthogonal mixture with at most $2^K K!$ components, see Definition 2.4.9.*

2.5.2 An Algorithm to Remove Label and Sign Switching

Examples 2.5.5 to 2.5.7 can be dealt with by applying an approach proposed by Stephens (2000) for clustering inference. Clustering inference assigns the observations in a given data set to different clusters and estimates the cluster-specific parameters from the data.[14] When formulated in terms of a decision-theoretic approach, the tasks involved in clustering inference, as estimating parameters or assigning an observation to one of the clusters, can be understood in an abstract way as choosing an action a from a set of actions $\mathcal{A}$ in such a way that the posterior expected loss, or expected risk, measured by a predefined loss function L_0, is minimized. For instance in Example 2.5.5, the set of possible actions in one step of the algorithm consists of applying any out of all possible column permutations to X_s. Similarly, in Example 2.5.6, the set of possible actions in the same step of the corresponding algorithm consists of changing any subset of the column signs of X_s.

The algorithm proposed by Stephens (2000) assigns each of N observations to one out of C clusters. The true distribution on C-group clusterings is denoted $P(\theta)$ and is an $N \times C$

[14] The number of clusters may be predefined, or may likewise be estimated from the data, e.g. using reversible jump MCMC methods, see Richardson and Green (1997) or by allowing for empty clusters.

matrix, and the estimated distribution on C-group clusterings of the observations among the clusters is denoted Q and is also an $N \times C$ matrix. Each row of $P(\theta)$ and Q, respectively, sums up to one, because every observation must be assigned to exactly one of the clusters. The proposed MCMC procedure minimizes the Monte Carlo risk as an approximation of the posterior expected loss, where the chosen loss function is the Kullback-Leibler divergence between $P(\theta)$ and Q, i.e.

$$L_0(Q, \theta) = \sum_{i=1}^{N} \sum_{j=1}^{C} p_{i,j}(\theta) \log \left(\frac{p_{i,j}(\theta)}{q_{i,j}} \right), \tag{2.41}$$

where $p_{i,j}(\theta)$ and $q_{i,j}$ denote the element in row i and column j of $P(\theta)$ and Q. The classification probabilities are calculated as

$$p_{i,j}(\theta) = \frac{\pi_j f(x_i; \phi_j, \eta)}{\sum_{l=1}^{C} \pi_l f(x_i; \phi_l, \eta)}, \tag{2.42}$$

where $f(\phi_j, \eta)$ denotes the distribution within cluster j, where ϕ_j are the parameters specific to cluster j and η are the parameters that are the same across all clusters, and x_i denotes the i^{th} observation.

The algorithm consists of two steps, first choosing $\hat{Q}$ such that the overall loss is minimized, i.e.

$$\hat{q}_{i,j} = \arg\min_{q_{i,j}} \sum_{s=1}^{S} \sum_{i=1}^{N} \sum_{j=1}^{C} p_{i,j}(\theta_s(\hat{D}_s)) \log \left(\frac{p_{i,j}(\theta_s(\hat{D}_s))}{q_{i,j}} \right), \tag{2.43}$$

where $\theta_s(\hat{D}_s)$ denotes the transformation of the parameters sampled in iteration s, θ_s, by the matrix $\hat{D}_s$, which is a permutation matrix. At initialization, $\hat{D}_s = I_K$ for all $s \in \{1, \ldots, S\}$.

The estimates for $q_{i,j}$ in the first step obtain as

$$\hat{q}_{i,j} = \frac{1}{S} \sum_{s=1}^{S} p_{i,j}(\theta_s(\hat{D}_s)), \tag{2.44}$$

see Stephens (2000).

In the second step, for every $s \in \{1, \ldots, S\}$, the permutation $\hat{D}_s$ is chosen such that the loss in each draw is minimized subject to the updated $\hat{Q}$, i.e.

$$\hat{D}_s = \arg\min_{D_s} \sum_{s=1}^{S} \sum_{i=1}^{N} \sum_{j=1}^{C} p_{i,j}(\theta_s(D_s)) \log \left(\frac{p_{i,j}(\theta_s(D_s))}{\hat{q}_{i,j}} \right). \tag{2.45}$$

Convergence is reached when in the second step, no further relabeling takes place.

This algorithm can directly be applied to Example 2.5.5, if the columns of X_s are considered as the observations, hence $N = K$, and there exist $C = K$ clusters. $f(\phi_j, \eta)$ is the joint distribution of every K^{th} element of $\text{vec}(X')$, whose distribution is given in Equation (2.34),

and which is hence also a normal distribution. ϕ_j contains the mean vector and covariance matrix of this normal distribution, η is empty, since there are no common parameters. The probabilities to assign an observation to one of the clusters, or the mixture proportions, can be assumed to be $\frac{1}{K}$ each, since every X_s contains exactly one representation of each column, but the labeling within each X_s is unknown. Thus the π_j in Equation (2.42) cancel out. The classification probability $p_{i,j}(\theta)$ is the probability of column i to fall into cluster j, where $\sum_{j=1}^{K} p_{i,j}(\theta) = 1$. If the distributions of the columns of X_s can be sufficiently well discriminated between, the resulting classification probability matrices for all X_s should therefore be close to I_K, which indicates that the columns are successfully relabeled.

Next, consider Example 2.5.6, where the columns of X_s undergo individual sign reflections. In this case, the columns of the X_s can be treated as K individual samples with $N = 1$ and $C = 2$ clusters for each sample. Each cluster is characterized by a normal distribution, where it is known that reversing the sign of the mean vector of the first cluster yields the mean vector of the second cluster, and their covariance matrices are identical. This information can be exploited in the algorithm, which is now different with respect to the orthogonal transformation used: Instead of relabeling, the signs of the column vectors of X_s are reversed. If the distribution of the column entries of X_s can be sufficiently well discriminated against its counterpart with reversed sign, the resulting classification probabilities should be close to 1 for one cluster, and close to zero for the other. As the effects of switching the column sign and assigning the column to the other cluster cancel out against each other, it is assumed for identification that the first of the two clusters is the one with the highest classification probability. A similar approach has been suggested for Bayesian confirmatory factor analysis by Erosheva and Curtis (2013).

Eventually, consider Example 2.5.7, where the columns X_s undergo a combination of permutations and sign switching. The columns of X_s are again considered as the observations, hence $N = K$, but, unlike in Example 2.5.5, there exist $C = 2K$ clusters now, two for each column, where the mean of the second is the mean of the first with reversed sign. To remove the orthogonal mixing in this case, a three-step algorithm is required. In the first step, $\hat{Q}$ is again chosen to minimize the overall loss. In the second step, the columns are relabeled as in Example 2.5.5. In the third step, the signs are switched as in Example 2.5.6. If the distributions of the column entries and their reflections can be sufficiently well discriminated between, K clusters should remain empty, and the matrix of assignment probabilities for the remaining clusters should again be close to the identity matrix.

If the algorithm is applied to the example discussed in Section 2.3.3 under the additional assumption that θ_1 and θ_2 follow bivariate normal distributions, $N = 2$ observations are assigned to $C = 2$ clusters, where each cluster must contain exactly one observation. In the first step, the mean vectors and covariance matrices for θ_1 and θ_2, which correspond to the mean vectors and covariance matrices for the cluster means, are estimated, and the classification probabilities $p_{i,j}(\theta)$ are estimated for each element of the sample. The average classification probabilities $q_{i,j}(\theta)$ are then estimated as in Equation (2.44), and in the second step, the labels of θ_1 and θ_2 are permuted wherever this reduces the Kullback-Leibler

divergence. The algorithm then returns to the first step. The fifth column of Figure 2.1 shows the resulting distribution of θ_1 in the top panel and that of θ_2 in the bottom panel. The true distributions, shown in the panels in the first column of the figure, are recovered very well by the relabeling algorithm.

2.6 Orthogonal Mixing Beyond Label and Sign Switching

After dealing with the special cases of orthogonal mixing involving only reflection and permutation matrices, where the orthogonal mixture distributions are finite in the sense of Definition 2.4.9, the general representation of orthogonal mixtures in Equation (2.38) is now considered, allowing D_s to be any orthogonal matrix. Since there are infinitely many distinct orthogonal matrices for $K > 1$, the resulting orthogonal mixture is infinite in the sense of Definition 2.4.10.

Consider an orthogonally mixed sample $\{X_s\}_{s=1}^S$, where X_s follows the distribution given in Equation (2.38). In the previous section, it was shown that if the orthogonal mixing is constrained to label or sign switching or combinations of both, the relabeling approach by Stephens (2000) can be applied to remove it. The illustrations in Examples 2.5.5 to 2.5.7 showed that these approaches do not require knowledge about the orthogonally invariant distribution.

2.6.1 An Algorithm to Remove Orthogonal Mixing

In the case of an infinite orthogonal mixture, relabeling and sign adjustment as in Examples 2.5.5 and 2.5.6 are no longer applicable, and the distribution of D is generally unknown. As a remedy, the orthogonal mixture is assumed to be finite with $C = S$ mixture components, each of which is characterized by a specific orthogonal matrix D_c. Since there is exactly one realization of every D_c in the sample, estimates are equivalent to estimates of the orthogonal matrices D_s for each element of the sample $\{X_s\}_{s=1}^S$. The classification probabilities from Equation (2.42) are therefore all $\frac{1}{S}$. To obtain estimates for $\tilde{M}$ and $\tilde{\tilde{\Sigma}}$, the X_s must be transformed such that they are all fall into the same cluster. To find the D_s that achieve this, the loss function must be changed accordingly.

Assume first that the orthogonally invariant distribution is known. The orthogonal mixing could then be removed by orthogonally transforming each X_s such that the Kullback-Leibler divergence between the orthogonally invariant distribution and the empirical distribution of the accordingly transformed draws is minimized, as in Stephens (2000). As Theorem 2.4.1 allows to express the required orthogonal matrix in terms of $K(K-1)/2$ angles $\{\gamma_h\}_{h=1}^{K(K-1)/2}$ and one reflection parameter r_K, such that a numerical optimization becomes feasible.

Next, consider the more realistic case that the orthogonally invariant distribution is not known. At least the first of the following two assumptions must hold in order for the proposed

approach to remove orthogonal mixing to work. The normality assumption is dropped here, only ellipticity of the orthogonally invariant distribution is required.

Assumption 2.6.1. *The underlying orthogonally invariant distribution in the sense of Definition 2.4.11 is elliptical, i.e.*

$$f(\tilde{X};\theta,I_K,1) \propto \det(\tilde{\bar{\Sigma}})^{-\frac{1}{2}} g((\mathrm{vec}(\tilde{X}')-\mathrm{vec}(\tilde{M}'))'\tilde{\bar{\Sigma}}^{-1}(\mathrm{vec}(\tilde{X}')-\mathrm{vec}(\tilde{M}'))), \tag{2.46}$$

for some function $g(\cdot)$, where $\tilde{\bar{\Sigma}}$ is again defined as in Equation (2.35), see e.g. Muirhead (1982), Definition 1.5.2.

This assumption ensures that under a quadratic loss function, the loss induced for an arbitrary orthogonal transformation of both $\tilde{X}$ and $\tilde{M}$ stays the same. As the row vectors of X_s by definition follow independent K-variate distributions, this implies that all these distributions are in turn also elliptical.

Assumption 2.6.2. *The mean $\tilde{M} = (\tilde{\mu}_1,\ldots,\tilde{\mu}_n)$ of the underlying orthogonally invariant distribution is known.*

If both assumptions hold, an easy to implement way to remove the orthogonal mixing and to find a $\hat{D}_s'$ that reverses the effect of the unknown D_s contained in X_s is to minimize the distance between X_s and $\tilde{M}$. For $\bar{\Sigma} = cI_{nK}$ and some $c \in \mathbb{R}_+$, the distance to be minimized is the Euclidean distance, so the problem to be solved corresponds to the orthogonal Procrustes problem, see e.g. Green (1952). The required $\hat{D}_s'$ then obtains as the solution of the corresponding minimization,

$$\hat{D}_s' = \underset{D}{\mathrm{argmin}}\,\mathrm{tr}((X_sD-\tilde{M})'(X_sD-\tilde{M})) \quad \text{subject to} \quad D \in O(K) \tag{2.47}$$

for each $s \in \{1,\ldots,S\}$.

Kristof (1964) and Schönemann (1966) provide a solution to the orthogonal Procrustes problem. A solution corresponding to that of Kristof (1964) was found independently by Roppert and Fischer (1965). While the initial approach required a time-consuming complete enumeration step, using a singular value decomposition allows for facilitate the algorithm to a large extent, see e.g. Golub and van Loan (2013). Appendix 2.B describes the orthogonal Procrustes algorithm in more detail, largely following Schönemann (1966).

2.6.2 The Weighted Orthogonal Procrustes Algorithm

In most cases, it cannot reasonably be assumed that $\bar{\Sigma} = cI_{nK}$ for some $c \in \mathbb{R}_+$, hence instead of the Euclidean distance, a different distance measure must be used to take the heteroskedasticity into account. Recall that by Definition 2.4.1, orthogonal transformations are length- and angle-preserving. Accordingly, the orthogonal Procrustes procedure in Equation (2.47) uses information from the lengths of the $x_{j,s}$ and the angles between $x_{j_1,s}$ and $x_{j_2,s}$ for $j_1 \neq j_2$, which is the exact same information as from the lengths of the $\tilde{x}_{j,s}$ and the angles between $\tilde{x}_{j_1,s}$ and $\tilde{x}_{j_2,s}$ for $j_1 \neq j_2$. If the $\mathrm{vec}(X_s')$ are scaled by an arbitrary weights

matrix, $W \in \mathbb{R}^{nK \times nK}$, this information is generally lost. This would thus be the case if $\bar{\Sigma}$ were known and $W = \bar{\Sigma}^{-1}$ was chosen, such that the distance to be minimized would correspond to the Mahalanobis distance. Conversely, if the weights matrix has the form $W = \text{diag}(w) \otimes I_K$, where $w \in \mathbb{R}^n$ and $\text{diag}(w)$ maps the vector w onto a conformably dimensioned diagonal matrix, the lengths of the vectors are changed, but the angles are preserved. It can be verified that

$$W^{-\frac{1}{2}} \text{vec}(X_s') = \text{vec}\left(X_s \text{diag}\,(w)^{-\frac{1}{2}}\right), \tag{2.48}$$

and hence, the weighted minimization changes accordingly, such that the estimate for D_s' obtains as

$$\hat{D}_s' = \underset{D}{\text{argmin}}\, \text{tr}((X_s D - \tilde{M})' \text{diag}(w)^{-1} (X_s D - \tilde{M})) \quad \text{subject to} \quad D \in O(K). \tag{2.49}$$

This modification of the orthogonal Procrustes problem is called weighted orthogonal Procrustes problem. The algorithm to obtain its solution proceeds accordingly to that for the (unweighted) orthogonal Procrustes problem, explained in Appendix 2.B. Its properties are discussed e.g. by Lissitz et al. (1976) and Koschat and Swayne (1991).

Therefore it is merely necessary to choose the vector of weights $w = (w_1, \ldots, w_n)$ appropriately. If $\bar{\Sigma}$ is known, or an estimate is at hand, the w_j for $j \in \{1, \ldots, n\}$ can be calculated as

$$w_j = \det\left(\Sigma_j\right)^{-\frac{1}{K}}, \tag{2.50}$$

where estimates for the Σ_j can be used where the actual values are unknown. Note that

$$\det\left(w_j \text{Cov}(x_j)\right) = \det\left(\Sigma_j^{-\frac{1}{K}} \text{Cov}(x_j)\right) = 1, \tag{2.51}$$

where the determinant is similarity-invariant and therefore remains unchanged for the covariance matrix of a sample in which all X_s are transformed by the same orthogonal matrix. In terms of this measure, the heteroskedasticity is thus dealt with by choosing the weights accordingly, where only the lengths of the $x_{j,s}$ are affected and no information about the angles $x_{j_1,s}$ and $x_{j_2,s}$ for $j_1 \neq j_2$ is lost. If no estimate for $\bar{\Sigma}$ is at hand, the orthogonally mixed sample $\{X_s\}_{s=1}^S$ can be used to calculate the weights as

$$w_j = S\left(\sum_{s=1}^{S} \sqrt{x_{j,s}' x_{j,s}}\right)^{-1}, \tag{2.52}$$

i.e. the inverse of the average length of the $x_{j,s}$. Note that orthogonal transformations are length-preserving, and hence, this measure takes the same value for $\{X_s\}_{s=1}^S$ as for $\{\tilde{X}_s\}_{s=1}^S$. Note that

$$E(w_j \|x_j\|) = 1, \tag{2.53}$$

and thus the average length of every $x_{j,s}$ is scaled to 1 for every $j \in \{1, \ldots, n\}$, accordingly taking care of the heteroskedasticity. Both measures are compared in the simulation study in Section 2.7.

Assumption 2.6.2 requires that the mean $\tilde{M}$ of the underlying orthogonally invariant distribution is known. This is, however, not a realistic assumption. It is possible, however, to develop an iterative approach along the lines of the expectation-maximization (EM) algorithm by Dempster et al. (1977), using the estimate $\hat{\tilde{M}}$ instead. The corresponding algorithm proceeds as follows, where certain distinctions apply, depending on which weighting scheme is used, where the unweighted orthogonal Procrustes corresponds to equal weighting. Recall that the first weighting scheme, using weights defined as in Equation (2.50), exploits information from $\bar{\Sigma}$ or its current estimate, and the second weighting scheme, using weights defined as in Equation (2.52), exploits information about the length of the $x_{j,s}$.

In the following, the subscript (τ) denotes the parameter estimates and weights after iteration τ of the algorithm. As throughout the rest of the chapter, only the final parameter estimates and weights at convergence are of interest, the subscript only appears here, in order to illustrate how the parameter estimate from the previous iteration $\hat{\tilde{M}}_{(\tau-1)}$ influences the subsequent iteration.

Algorithm 2.6.1.

1. *For every X_s, solve the (weighted) orthogonal Procrustes problem conditional on $\hat{\tilde{M}}_{(\tau-1)}$ and $w_{(\tau-1)}$, where $w_{(\tau-1)} = 1_n$ for the unweighted orthogonal Procrustes problem, to obtain a sequence of orthogonal matrices $\{\hat{D}_{s(\tau)}\}_{s=1}^{S}$.*
2. *Update $\hat{\tilde{M}}_{(\tau)} = \frac{1}{S}\sum_{s=1}^{S} X_s \hat{D}'_{s(\tau)}$.*
3. *Update $\hat{\tilde{\Sigma}}_{j(\tau)} = \frac{1}{S}\sum_{s=1}^{S}(\hat{D}_{s(\tau)}x_{j,s} - \hat{\tilde{\mu}}_{j(\tau)})(\hat{D}_{s(\tau)}x_{j,s} - \hat{\tilde{\mu}}_{j(\tau)})'$, if the first weighting scheme is used.*
4. *Update $w_{(\tau)} = (w_{1(\tau)}, \ldots, w_{n(\tau)})$, where $w_{j(\tau)} = \det\left(\hat{\tilde{\Sigma}}_{j(\tau)}\right)^{-\frac{1}{K}}$, if the first weighting scheme is used.*
5. *If the difference between $\hat{\tilde{M}}_{(\tau-1)}$ and $\hat{\tilde{M}}_{(\tau)}$ is sufficiently large, proceed with step 1.*

Algorithm 2.6.1 requires an initialization $\hat{\tilde{M}}_{(0)}$ and, if the first weighting scheme is used, also an initialization $w_{(0)}$. For convenience, $\hat{\tilde{M}}_{(0)} = X_S$ is chosen, i.e. the last draw in the orthogonally mixed sample, and the initial weights $w_{(0)}$ are calculated according to the second weighting scheme. If the second weighting scheme is chosen, Steps 3 and 4 are not required, as the weights are not iteratively updated, but stay the same, since the lengths of the $x_{j,s}$ stay the same under orthogonal transformations. Hence, the second weighting scheme does not exploit new information due to changes in $\{\hat{D}_{s(\tau)}\}_{s=1}^{S}$.

Convergence is assumed in iteration τ for

$$\left\| \operatorname{vec}\left(\hat{\tilde{M}}_{(\tau)} - \hat{\tilde{M}}_{(\tau-1)}\right)\right\|^2 \leq \omega, \tag{2.54}$$

where ω denotes a threshold value, which is generally set to 10^{-9} in the simulation study in Section 2.7. The case where Assumption 2.6.2 holds is nested in the algorithm, implying that $\hat{\tilde{M}}_{(0)} = \tilde{M}$ and the algorithm stops after the first iteration.

As there are infinitely many ways to parameterize the underlying orthogonally invariant distributions by applying any orthogonal matrix D to obtain $\theta(D)$, the algorithm converges to the mean of some representation of the original orthogonally invariant distribution, which depends on its initialization. It must be noted, however, that the effect of the initialization can be reversed by an appropriate orthogonal transformation of $\hat{\tilde{M}}$, see Section 3.3. This can also be understood from the fact that the solution of the orthogonal Procrustes problem in Equation (2.49) is the same if both X_s and $\tilde{M}$ are transformed by the same orthogonal matrix. As the solution is unique for X_s conditional on $\tilde{M}$, see e.g. Schönemann (1966) or Lissitz et al. (1976), using for the initialization some $\hat{\tilde{M}}_{(0)}D$, where $D \in O(K)$, accordingly yields $\hat{\tilde{M}}_{(\tau)}D$ at convergence. For more details on the convergence properties of the algorithm, discussed in the context of Bayesian estimation of latent dynamic factor models, see Chapter 3.

2.7 Simulation Study

In the following, I analyze the behavior of Algorithm 2.6.1 for equal weights as well as the two weighting schemes discussed in Section 2.6. To this end, I simulate orthogonally mixed samples $\{X_s\}_{s=1}^{S}$ with $S = 10,000$ throughout, for $n \in \{1, \dots, 50\}$ and $K \in \{2, \dots, 6\}$, taking different distributions into account, which all satisfy Assumption 2.6.1. The considered distributions are the normal distribution and the Student t distribution with 3, 10 and 25 degrees of freedom, respectively. The required ellipticity property also holds e.g. for other symmetric α-stable distributions and Laplace distributions, which are not discussed here for brevity. Note that Algorithm 2.6.1 proceeds in a purely deterministic way as it processes the orthogonally mixed samples, hence the results obtained for the same orthogonally mixed sample are perfectly reproducible.

The mean parameters are all independently drawn from a uniform distribution over the interval $[4; 6]$ hence $E[\tilde{\mu}_{j,k}] = 5$ for all $j \in \{1, \dots, n\}$ and $k \in \{1, \dots, K\}$, and the covariance matrices are all independently drawn from Inverse Wishart distributions with the scale parameter chosen either as $\Psi = 0.02I_K$ or as $\Psi = 0.2I_K$ and the degrees of freedom $\nu = 10$. The first choice of Ψ, referred to in the following as the *small variance* setting, results in $E[\tilde{\tilde{\sigma}}^2_{(j-1)K+k}] = 0.2$, and the second, referred to as the *large variance* setting, results in $E[\tilde{\tilde{\sigma}}^2_{(j-1)K+k}] = 2$ for all $j \in \{1, \dots, n\}$ and $k \in \{1, \dots, K\}$.

The orthogonally mixed samples are generated as follows: First, a sample from the underlying orthogonally invariant distribution $\{\tilde{X}_s\}_{s=1}^{S}$ is drawn from the respective normal or Student t distributions. Next, for each s, a set of angles $\{\gamma_h\}_{h=1}^{\frac{K(K-1)}{2}}$ is drawn from a uniform distribution with range $(-\pi, \pi]$, and a reflection parameter $r_{K,K}$ is drawn from a Rademacher distribution, i.e. $r_{K,K}$ takes values -1 or $+1$ with probability $\frac{1}{2}$ each. Using the result from Theorem 2.4.1, this allows to construct a sample of random orthogonal matrices $\{D_s\}_{s=1}^{S}$, which are then

used to postmultiply the $\{\tilde{X}_s\}_{s=1}^S$ by, in order to obtain the orthogonally mixed sample as $\{X_s\}_{s=1}^S = \{\tilde{X}_s D_s\}_{s=1}^S$.

When applying Algorithm 2.6.1 to these orthogonally mixed samples, all configurations with both settings are processed by the orthogonal Procrustes algorithm with equal weights, weights according to the first and according to the second weighting scheme. Eventually, the algorithm is checked for the case that Assumption 2.6.2 holds, where only one iteration of the algorithm is required, and for the case where Assumption 2.6.2 does not hold, and hence, the algorithm iterates until convergence.

2.7.1 Required Number of Iterations Until Convergence

First, I report the number of iterations required by Algorithm 2.6.1 to reach convergence. Tables 2.1 and 2.2 show the results for all values of n and K for small and large variances, respectively, for normally distributed data. It can be seen that for both weighting schemes, values of $n \geq 10$ always result in the algorithm converging after three iterations in the small variance case, and after three or four iterations, in very few cases for small n and large K, up to five iterations, in the large variance case. For values of $n < 10$, there are several cases where under both weighting schemes, the algorithm requires up to 26 or 21 iterations until convergence in the large variance case. The equal weighting requires substantially more iterations until convergence, up to 34 for the small variance case, and up to 86 for the large variance case. Moreover, with increasing K, the number of required iterations generally also increases, and the number of iterations for $n < 20$ is frequently in the double digits.

Tables 2.3 and 2.4 show the results for Student t distributed data with 3 degrees of freedom. For the small-variance data, both weighting schemes yield a number of required iterations that barely exceeds five if $n \geq 10$. It reaches up to 15 and 13, respectively, for smaller values of n. For the equal weighting, the number of required iterations again increases with K and is generally substantially larger, frequently exceeding ten for $n \leq 20$, with three cases where no convergence is reached within 1,000 iterations. For the large-variance data, both weighting schemes yield a required number of iterations that does not exceed five for $n \geq 14$, whereas it reaches up to 29 and 26, respectively, for smaller values of n. The equal weighting requires up to 20 iterations for $n \geq 15$, and much more for smaller values of n. In two cases, no convergence is reached within 1,000 iterations. For Student t distributed data with 10 degrees of freedom, where the results are shown in Tables 2.5 and 2.6, respectively, the number of iterations required until convergence for the small-variance case generally does not exceed four for $n \geq 10$, while it reaches up to 39 and 43, respectively, for small values of n. Convergence under the equal weighting takes much longer again, frequently requiring to 50 iterations, some cases even exceeding that and one case with no convergence within 1,000 iterations. Results for the large-variance case are similar. Tables 2.7 and 2.8 show the results for Student t distributed data with 25 degrees of freedom, which are similar to the previous case.

Altogether, for the Student t and normally distributed data, the equal weighting provides reasonably quick convergence for $n \geq 20$, while the two weighting schemes converge much faster even for $n \geq 10$.

2.7.2 Root Mean-Squared Errors for the Mean Estimates

Next, consider the root mean-squared errors (RMSE) for the $\hat{\tilde{M}}$ obtained from Algorithm 2.6.1 using either equal weighting or any of the two weighting schemes and applying the algorithm to data sets following the aforementioned distributions with small and large variances, respectively. The RMSE is estimated as

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tilde{M}}_r - \tilde{M}\right)^2}, \tag{2.55}$$

where R denotes the number of samples used for every scenario. For comparison, the algorithm is either initialized with the true $\tilde{M}$ or with the last element from the sample X_S. The former case is not a realistic scenario, however, assuming that the parameter to be estimated is known beforehand. It does, however, provide the best possible estimates and hence a benchmark for the estimates obtained under in the latter case. If $\tilde{M}$ is known, no updating of $\hat{\tilde{M}}$ is necessary and therefore, only one iteration of the algorithm is required.[15] As $\hat{\tilde{M}}$ does not converge to the $\tilde{M}$, but an orthogonal transformation thereof, it is orthogonally transformed with respect to $\tilde{M}$ before plugging it into Equation (2.55). For all scenarios, $R = 50$ is chosen and the averages of the RMSE over all elements of $\tilde{M}$ are reported.

Table 2.9 shows the results for the normally distributed data. Differences between the equal weighting and the two weighting schemes are negligible, and overall, the average RMSE is about five to eight times bigger for large compared to small variances. Increasing n tends to reduce the average RMSE, while increasing K tends to lead to an increase in the average RMSE. This holds throughout for $n >> K$, while results for $n \approx K$ or even $n < K$ generally show a much larger average RMSE. The results for unknown and known $\tilde{M}$ are similar, which shows that Algorithm 2.6.1 provides a good estimate for $\tilde{M}$ from the orthogonally mixed sample.

Tables 2.10 to 2.12 show the results for the Student t distributed data. The differences in average RMSE between the small and large variance cases ar rather small compared to the case of normally distributed data. Throughout, the RMSE decreases in n and increases in K. The average RMSE for cases where $n \approx K$ or even $n < K$ is substantially higher than for other cases. Differences in average RMSE between equal weighting and the two weighting schemes are rather small, and the similarity between results using $\tilde{M}$ in Step 1 of the algorithm and those where $\tilde{M}$ is unknown indicate that the algorithm does not require any knowledge about the underlying orthogonally invariant distribution. This holds overall,

[15] Since the first weighting scheme is identical to the second in the first iteration, results for both weighting schemes are identical in this case.

so altogether, Algorithm 2.6.1 satisfies the requirement formulated at the beginning of Section 2.6 in analogy to the relabeling approach by Stephens (2000) for the sign and label switching, where the average RMSE generally increases as K increases and decreases as n increases.

2.8 Empirical Distributions after Postprocessing

Having found that overall, the mean of the underlying orthogonally invariant distribution can be estimated from the orthogonally mixed samples by Algorithm 2.6.1, I now investigate whether the same holds for the distribution as a whole. For brevity, I only consider the normally distributed and Student t distributed data with $\nu = 3$ degrees of freedom, and the large variance case.

2.8.1 Evidence from Quantile-Quantile Plots

Figure 2.2 shows quantile-quantile (QQ) plots for the percentiles of one randomly chosen x_{jk} for the normally distributed data and large variances for $n = 10$ and $n = 50$, and $K = 2$ and $K = 6$, respectively. On the horizontal axis, each plot shows the percentiles of the originally simulated data without orthogonal mixing. In each row, the first plot shows the percentiles for the orthogonally mixed data on the vertical axis, while the three following panels show the according percentiles for the data postprocessed by Algorithm 2.6.1 on the vertical axis, with equal weighting and the above described weighting schemes. For $n = 10$ and $K = 2$, the central percentiles are recovered very well, while the extreme percentiles are biased towards the center of the distribution. This is least pronounced for the first weighting scheme. The same can be observed for $n = 10$ and $K = 6$. For $n = 50$ and $K = 2$ as well as for $n = 50$ and $K = 6$, the weighting is irrelevant, and the restored sample resembles the original one even in the extreme percentiles. Figure 2.3 shows the QQ plots for Student t distributed data with $\nu = 3$ degrees of freedom. In all considered cases, even the extreme percentiles are recovered very well by the algorithm. Overall, for the normal and Student t distributed data, especially for small values of n and large values of K, the QQ plots are slightly tilted to the right, which indicates that the algorithm tends to produce a distribution whose lower percentiles are overestimated and whose upper percentiles are underestimated.

Table 2.13 shows the average differences between the simulated and the restored data for all nK elements of X_s for normally distributed data, where the 5%, 25%, 50%, 75% and 95% quantiles are reported. Keeping in mind that the parameter values are centered around 5, the differences are overall very small. Regarding the median, the difference between the simulated and restored data quantiles is zero on average. The observation from the QQ plots that the values for the lower quantiles tend to be too high and those for the upper quantiles tend to be too low is confirmed by the results. The differences are larger for the outer quantiles, and they increase as K increases, while they decrease as n increases. The same holds for the standard errors of the differences. Table 2.14 shows the average differences for the Student

t distributed data with $\nu = 3$ degrees of freedom. The results generally resemble those for normally distributed data, except for slightly larger average differences and corresponding standard errors. This indicates that Algorithm 2.6.1 is generally able to restore the properties of this distribution from the orthogonally mixed sample.

2.8.2 Effects of Postprocessing on Low and High Quantiles

Next, I take a closer look at the algorithm's tendency to overestimate the low quantiles of the distribution and underestimate the high quantiles of the distribution. This indicates that Algorithm 2.6.1 in fact removes more variation from the orthogonally mixed sample than it should, including such variation that occurs in the sample from the underlying orthogonally invariant distribution. It is hence overfitting with regard to the center of the distribution and neglecting the tails, about which little information is at hand. Similar problems with recovery of the tails of the distributions of interest have been observed for discrete relabeling e.g. by Sperrin et al. (2010), who suggest a stochastic relabeling approach as a remedy. Finding a similar approach for the proposed algorithm is beyond the scope of this chapter, however in the following, I investigate the extent of the distortion.

To outline the problem, note that two points X_{s_1} and $X_{s_2} = X_{s_1}D$, where $D \in O(K)$ the algorithm yields $\hat{D}_{s_1} = \hat{D}_{s_2}D$, and hence maps both points onto the same point, so $\tilde{X}_{s_1} = \tilde{X}_{s_2}$. Accordingly, for two points X_{s_1} and $X_{s_2} = cX_{s_1}D$, the algorithm transforms both points in such a way that $c\tilde{X}_{s_2} = \tilde{X}_{s_1}$. Therefore, for $n = 1$, all vectors $x_{1,s}$ for $s \in \{1, \dots, S\}$ are projected onto each other, which is illustrated for $K = 2$ in Figure 2.4. Note that the space that all $x_{1,s}$ are mapped into is an orthogonal transformation of the $\mathbb{R}_+$. Accordingly, for $n < K$, Algorithm 2.6.1 maps the points from the $\mathbb{R}^K$ into an orthogonal transformation of the the $\mathbb{R}^n_+$.

For the case shown in Figure 2.4, the variance in the postprocessed sample is much smaller than in the sample from the orthogonally invariant distribution. This is in line with the previously made observation that the extreme quantiles are biased towards the center of the distribution. Figures 2.5 and 2.6 show the cases $n = 2$ and $n = 3$, where the red, blue and green point clouds denote the vectors $x_{1,s}$, $x_{2,s}$ and $x_{3,s}$ for $s \in \{1, \dots, S\}$. Decompose the draws from the orthogonal mixture distribution as

$$X_s = (\tilde{M} + E_s)D_s, \tag{2.56}$$

and consider the orthogonal Procrustes decomposition from Equation (2.47), which yields the $\hat{D}'_s$ that minimizes

$$\operatorname{tr}(\hat{E}'_s\hat{E}_s) = \operatorname{tr}((X_s\hat{D}'_s - \tilde{M})'(X_s\hat{D}'_s - \tilde{M})) \quad \text{subject to} \quad \hat{D}'_s \in O(K). \tag{2.57}$$

This expression can be rewritten as

$$\operatorname{tr}(\hat{E}'_s\hat{E}_s) = \operatorname{tr}(((\tilde{M} + E_s)D_s\hat{D}'_s - \tilde{M})'((\tilde{M} + E_s)D_s\hat{D}'_s - \tilde{M}))$$

$$
\begin{aligned}
&= \operatorname{tr}\left(((\tilde{M}+E_s)(D_s\hat{D}_s'-I_K)-E_s)'((\tilde{M}+E_s)(D_s\hat{D}_s'-I_K)-E_s)\right) \\
&= \underbrace{\operatorname{tr}\left(E_s'E_s\right)}_{\alpha_1} + \underbrace{2\operatorname{tr}\left(E_s'((\tilde{M}+E_s)(D_s\hat{D}_s'-I_K))\right)}_{\alpha_2} \\
&+ \underbrace{\operatorname{tr}\left(((\tilde{M}+E_s)(D_s\hat{D}_s'-I_K))'((\tilde{M}+E_s)(D_s\hat{D}_s'-I_K))\right)}_{\alpha_3}, \qquad (2.58)
\end{aligned}
$$

where E_s is the $n \times K$ matrix of errors under the orthogonally invariant distribution. Now consider the case

$$\operatorname{tr}(\hat{E}_s'\hat{E}_s) > \operatorname{tr}(E_s'E_s). \qquad (2.59)$$

This case can never occur, since $\hat{D}_s'$ in Equation (2.57) could be chosen as D_s', in which case $\alpha_2 = \alpha_3 = 0$. So

$$\alpha_1 + \alpha_2 + \alpha_3 \leq \alpha_1, \qquad (2.60)$$

i.e. α_1 is an upper bound for $\operatorname{tr}(\hat{E}_s'\hat{E}_s)$. This implies that

$$\alpha_2 + \alpha_3 \leq 0, \qquad (2.61)$$

and thus

$$-\alpha_2 \geq \alpha_3, \qquad (2.62)$$

where α_3 is the trace of a quadratic form and must therefore be positive. Consequently, if $\tilde{M}$ is known or can be consistently estimated,

$$\frac{\operatorname{tr}(\hat{E}_s'\hat{E}_s)}{\operatorname{tr}(E_s'E_s)} \leq 1. \qquad (2.63)$$

In the following, I run Algorithm 2.6.1 for orthogonally mixed samples from different distributions, where $S = 10,000$ and n and K vary, and look at the properties of the error ratio in Equation (2.63). Figures 2.7 and 2.8 show the quantiles for n up to 50 and K up to 7. As the plots look virtually identical, there is strong evidence that the algorithm actually converges to $\tilde{M}$ or an orthogonal transformation thereof. The upper bound of the error ratio can clearly be seen, and its median converges to 1 rather quickly as n increases, though the convergence speed is different for different values of K. For increasing K, the distance between the outer quantiles also seem to be decrease. Figures 2.9 and Figures 2.10 show the experiment's result if Student t distributed data is used. Again, there is no noticeable difference between the case where $\tilde{M}$ is known and where it is unknown. With regard to the behavior for increasing n and K, the results are similar as for the normally distributed data.

Next, I take a short look at what the weighting schemes do. Figures 2.11 and 2.12 show the outcomes under the first and second weighting schemes, respectively, where $\tilde{M}$ is unknown. Note that

$$\operatorname{tr}(\hat{E}_s' W \hat{E}_s) \leq \operatorname{tr}(E_s' E_s) \tag{2.64}$$

will not necessarily hold, so the upper bound also no longer holds. A look at the yellow median line, however, indicates that under the two weighting schemes, the median error ratio is generally closer to 1 than under the equal weighting. Especially for small n, the first scheme fares better than the second.

Eventually, I look at the mean error ratio from Equation (2.63) and its behavior as n and K increase. Figure 2.13 shows the according results for different K, additionally showing the mean plus and minus two standard deviations. The solid lines denote the case where $\tilde{M}$ is known, the dotted lines denote the case where it is unknown. The two standard deviation bands around the mean get narrower as n increases and are generally narrower for increasing K, which corresponds to the findings for the quantiles. Trying to find a way to estimate the error ratio in terms of n and K, the results indicate that

$$E\left(\frac{\operatorname{tr}(\hat{E}_s' \hat{E}_s)}{\operatorname{tr}(E_s' E_s)}\right) = 1 - \frac{K-1}{2n}, \tag{2.65}$$

which is what the red line in the plot shows. The fit is overall very good except for very small values of n. Figure 2.14 shows the results for Student t distributed data with $\nu = 3$ degrees of freedom. The results in this case are similar to those for the normally distributed data.

Figures 2.15 and 2.16 show the results under the two weighting schemes. The two standard deviation bands get narrower for increasing n and increasing K, and the dotted and solid lines are overall the same for $n > 10$, or $n > 20$ in the case of $K = 7$. Note, however, that the approximation of the error ratio for the equal weighting scheme is generally exceeded by the actual mean error ratio, which thus tends to be closer to one. Hence the weighting schemes provide better estimates of the errors under the orthogonally invariant distribution and thus the shape of the orthogonally invariant distribution from the orthogonally mixed sample.

2.9 Orthogonal Mixing and Factor Models

In the following, I consider the Gibbs sampling scheme for the static factor model described in Section 2.3, with and without the PLT identification constraint. The simulated example is the same as in Chapter 3, but is considered here with the aim of making the orthogonal mixing visible. As Algorithm 2.6.1 provides a tool to remove all orthogonal mixing from a sample, it is therefore possible to postprocess the Gibbs output in such a way that the orthogonal mixing is removed. Both the factors and the factor loadings are affected by the orthogonal mixing, however, it suffices to apply the algorithm to the factor loadings to evaluate the extent of

orthogonal mixing present in the output of a Gibbs sampler, measured as the variation of the resulting $\hat{D}_s$.

The loadings matrix that is used for simulating the data is

$$\Lambda = \begin{pmatrix} 0.100 & -0.200 & 0.500 & 0.600 & 0.100 & 0.174 & -0.153 & -0.470 & 0.186 & -0.577 \\ 0.000 & 0.200 & -0.100 & 0.400 & -0.900 & 0.429 & -0.392 & 0.652 & 0.282 & -0.541 \end{pmatrix}' \tag{2.66}$$

and the corresponding matrix of idiosyncratic covariances is diagonal, which corresponds to the *Frisch case* after Frisch (1934) by Scherrer and Deistler (1998), so

$$\Sigma = \text{diag}(0.990, 0.920, 0.740, 0.480, 0.180, 0.786, 0.823, 0.354, 0.886, 0.374). \tag{2.67}$$

The factors f_t and errors e_t for $t \in \{1, \dots, 100\}$ are jointly drawn from a normal distribution, using $\Omega = I_K$ and Σ, hence all factors and errors are mutually orthogonal. The prior hyperparameters for Λ are chosen as $\mu_{\lambda_i} = 0_K$ and $\Sigma_{\lambda_i} = I_K$, the ones for Σ are chosen as $\underline{\alpha}_i = \underline{\beta}_i = 1$ for all $i \in \{1, \dots, N\}$.

Next, I consider the data simulated using Λ and Σ under three different orderings, as in Chapter 3. Note that due to the different orderings, the effect of the constraints induced by the Dirac Delta prior and the truncation below at zero varies, i.e. even though all constraints lead to the same exact model identification, the information from the likelihood that enters the full conditional distribution of Λ is different, because different subsets of the information stored in the likelihood are are canceled out by the highly informative priors. If the orthogonally invariant priors are used, no constraints are imposed and the rotation problem is thus not solved, the full conditional distribution of Λ still contains all information from the likelihood, which comes at the price that the sampler output is orthogonally mixed. The aim of imposing the constraints via informative priors is consequently to suppress the orthogonal mixing in the best possible way without losing too much information from the likelihood.

The first ordering is the original one, thus the first two variables are chosen as factor founders, the second ordering chooses the second and third variables as factor founders, and the third ordering chooses the fifth and second variables. Note that the ordering among the factor founders matters due to the triangular form of the informative prior: The first factor founder is therefore the variable from which most of the likelihood information is overridden by the prior.

Both samplers, with and without constraints are run for $40,000$ iterations, of which the first half is discarded as burn-in. Being only interested in the factor loadings, I therefore consider the Gibbs sequence $\{\Lambda^{(z)}\}_{z=1}^{Z}$, where $Z = 20,000$. The orthogonal mixing is removed from this output by applying Algorithm 2.6.1. This yields a sequence of orthogonal matrices $\{\hat{D}^{(z)}\}_{z=1}^{Z}$, which are then decomposed according to Theorem 2.4.1. As $K = 2$, this decomposition gives an angle parameter $\hat{\gamma}^{(z)}$ and a reflection parameter $\hat{r}^{(z)} = \det(\hat{D}^{(z)})$ for every $z \in \{1, \dots, Z\}$, where $\hat{r}^{(z)} \in \{-1, 1\}$. Recall that the matrices $\{\hat{D}^{(z)}\}_{z=1}^{Z}$ depend on the initialization of the

algorithm and are merely an orthogonal transformation of the matrices $\{D^{(z)}\}_{z=1}^{Z}$, see Section 2.6. Therefore, it is possible to transform the $\hat{\gamma}^{(z)}$ into

$$\tilde{\hat{\gamma}}^{(z)} = \text{mod}(\hat{\gamma}^{(z)} + \delta, 2\pi) - \pi \tag{2.68}$$

for some $\delta \in \mathbb{R}$, or to reverse the signs of all $\hat{r}^{(z)}$, or both. The initialization of the algorithm with $\Lambda^{(Z)}$ yields $\hat{r}^{(Z)} = 1$ and $\hat{\gamma}^{(Z)} \approx 0$, so the resulting $\hat{\gamma}^{(z)}$ and $\hat{r}^{(z)}$ can approximately be understood as angles and reflections relative to the last draw of the sampler $\Lambda^{(Z)}$.

Figure 2.17 shows the according results for the first ordering of the variables, where the first plot shows the angles from the $\hat{D}^{(z)}$ without any constraints imposed on Λ and the rotation problem thus unsolved. The red line at the top of the plot indicates that $r^{(z)} = 1$ for all $z \in \{1, \ldots, Z\}$, so none of the orthogonal matrices has a negative determinant. In the case of $K = 2$, a negative determinant would imply that one factor switches its sign inbetween two iterations of the Gibbs sampler. Both factors switching their sign at once, however, could be diagnosed by a jump in $\hat{\gamma}$ of approximately $\pm\pi$.[16] None of this can be observed, instead, the sequence of $\hat{\gamma}$ looks like a random walk on the circle at first glance. As Algorithm 2.6.1 removes all orthogonal mixing from $\{\Lambda^{(z)}\}_{z=1}^{Z}$, it is not surprising that in the second plot, $\hat{\gamma}^{(z)} = 0$ and $\hat{r}^{(z)} = 1$ for all $z \in \{1, \ldots, Z\}$, which is precisely the property of a sample without any orthogonal mixing remaining. The third plot shows the accordingly transformed output for the constrained Gibbs sampler, which attempts to solve the rotation problem by imposing constraints on the $\Lambda^{(z)}$ via informative priors. The orthogonal mixing is far less pronounced than for the unconstrained sampler, but it is still clearly visible. There are no indications of sign switching, however, within the 20,000 iterations after the burn-in, the sampler reaches almost every point on the circle with respect to $\hat{\gamma}^{(z)}$.

Figure 2.18 shows the results for the second ordering of the variables. The first plot for the sampler without constraints looks similar to the first plot in the previous figure, and the second plot looks identical to the second plot in the previous figure, as here likewise, all orthogonal mixing has been removed from the sampler's output. The third plot, however, shows substantially less orthogonal mixing, with the values of $\hat{\gamma}^{(z)}$ almost exclusively ranging between -1.5 and 1. The corresponding choice of factor founders is thus more effective than the previous one in suppressing orthogonal mixing.

Figure 2.19 shows the results for the third ordering of the variables. Again, the first plot resembles the first plots in the previous two cases, and the second plot is identical to the second plot in these cases. The third plot, however, looks almost identical to the second, which indicates that here, the constraints successfully suppress the orthogonal mixing almost perfectly and are thus almost perfectly effective. This corresponds to an almost ideal choice of factor founders.

[16] Note that for $\gamma = \pi$, the corresponding rotation matrix is $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$.

2.10 Conclusion

In the present chapter, I introduce the concept of orthogonal mixing. An orthogonal mixture distribution is a mixture over orthogonally transformed versions of the same distribution. The scope of the chapter is constrained to elliptical distributions, where orthogonal mixing is relatively easy to handle. In finite samples from an orthogonal mixture distribution, the orthogonal transformations of each sample element can be expressed in terms of a unique orthogonal matrix. If these orthogonal matrices are constrained to permutation and reflection matrices, the corresponding orthogonal mixture is finite and the orthogonal mixing in the sample is label and sign switching, which are well-known phenomena from the Markov switching and mixture model literature, and for which relabeling algorithms have been suggested e.g. by Stephens (2000) and Frühwirth-Schnatter (2001). Generalizing the concept to infinite mixtures and allowing for all types of orthogonal matrices requires a different approach to remove the orthogonal mixing, for which I propose Algorithm 2.6.1, which removes the orthogonal mixing from each element of the sample with respect to the current estimate of the orthogonally invariant sample mean by solving the (weighted) orthogonal Procrustes problem and then updates the estimate of the orthogonally invariant sample mean. A simulation study shows that this algorithm works well for the considered elliptical distributions. Looking at the quantiles of the postprocessed samples indicates that the results are more condensed than they should be. I investigate on this issue, finding that the algorithm removes some variation beyond the variation present in the sample due to orthogonal mixing. The loss of variation, however, quickly goes to zero as the number of variables in the sample increases. Eventually, I consider a static factor model, for which I investigate the outcome of two Gibbs sampling approaches, where the first imposes constraints for model identification and the second does not. Algorithm 2.6.1 removes the entire orthogonal mixing from the second output and allows for a comparison of the extent of orthogonal mixing in the first output that obtains under different constraints that ensure exact identification of the factor model.

Tables

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	13	3	3	3	3	3	3	3	3	3	3	3	3	2	3
3	7	4	3	3	3	8	3	3	3	3	9	3	3	3	3
4	8	8	9	3	3	3	4	3	3	3	3	5	3	3	3
5	5	20	11	21	3	3	3	6	3	3	3	3	5	3	3
6	5	16	12	18	34	3	3	3	3	3	3	3	3	3	3
7	5	9	17	13	31	3	3	3	3	3	3	3	3	3	3
8	4	7	14	18	26	3	3	3	3	3	3	3	3	3	3
9	5	6	14	17	17	3	3	3	3	3	3	3	3	3	3
10	4	7	10	12	18	3	3	3	3	3	3	3	3	3	3
11	5	6	9	9	15	3	3	3	3	3	3	3	3	3	3
12	4	5	8	16	31	3	3	3	3	3	3	3	3	3	3
13	5	5	7	9	12	3	3	3	3	3	3	3	3	3	3
14	4	6	7	12	30	3	3	3	3	3	3	3	3	3	3
15	4	6	7	8	10	3	3	3	3	3	3	3	3	3	3
16	4	5	7	7	11	3	3	3	3	3	3	3	3	3	3
17	4	5	8	7	9	3	3	3	3	3	3	3	3	3	3
18	4	5	6	7	12	3	3	3	3	3	3	3	3	3	3
19	4	4	6	6	8	3	3	3	3	3	3	3	3	3	3
20	4	5	6	10	7	3	3	3	3	3	3	3	3	3	3
21	4	5	6	6	8	3	3	3	3	3	3	3	3	3	3
22	4	4	5	6	9	3	3	3	3	3	3	3	3	3	3
23	4	4	5	6	7	3	3	3	3	3	3	3	3	3	3
24	4	5	5	6	7	3	3	3	3	3	3	3	3	3	3
25	4	5	6	7	6	3	3	3	3	3	3	3	3	3	3
26	4	5	6	5	6	3	3	3	3	3	3	3	3	3	3
27	3	4	5	6	7	3	3	3	3	3	3	3	3	3	3
28	4	5	5	5	7	3	3	3	3	3	3	3	3	3	3
29	4	5	5	6	6	3	3	3	3	3	3	3	3	3	3
30	4	4	5	6	6	3	3	3	3	3	3	3	3	3	3
31	4	5	5	5	6	3	3	3	3	3	3	3	3	3	3
32	4	4	5	5	6	3	3	3	3	3	3	3	3	3	3
33	4	5	5	5	6	3	3	3	3	3	3	3	3	3	3
34	4	5	5	5	6	3	3	3	3	3	3	3	3	3	3
35	3	4	4	5	6	3	3	3	3	3	3	3	3	3	3
36	4	4	5	5	6	3	3	3	3	3	3	3	3	3	3
37	4	4	5	5	5	3	3	3	3	3	3	3	3	3	3
38	4	4	5	5	6	3	3	3	3	3	3	3	3	3	3
39	4	4	5	5	6	3	3	3	3	3	3	3	3	3	3
40	4	4	4	6	6	3	3	3	3	3	3	3	3	3	3
41	4	4	5	5	6	3	3	3	3	3	3	3	3	3	3
42	3	4	5	5	6	3	3	3	3	3	3	3	3	3	3
43	4	4	5	5	6	3	3	3	3	3	3	3	3	3	3
44	3	4	5	5	6	3	3	3	3	3	3	3	3	3	3
45	3	4	4	5	6	3	3	3	3	3	3	3	3	3	3
46	3	4	4	5	5	3	3	3	3	3	3	3	3	3	3
47	3	4	5	5	5	3	3	3	3	3	3	3	3	3	3
48	3	4	4	5	5	3	3	3	3	3	3	3	3	3	3
49	3	4	4	5	5	3	3	3	3	3	3	3	3	3	3
50	4	4	4	5	5	3	3	3	3	3	3	3	3	3	3

Table 2.1: Number of iterations for normally distributed data with small variances.

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	22	4	4	3	3	4	4	4	3	3	3	4	4	3	3
3	28	12	4	4	4	4	4	4	4	4	4	4	4	4	3
4	11	33	19	4	4	4	26	4	4	4	4	21	4	4	4
5	12	19	22	40	86	4	4	4	5	4	4	4	4	5	4
6	11	28	12	60	45	4	4	4	5	7	4	4	4	5	7
7	7	24	33	35	34	4	4	4	10	11	3	4	4	10	10
8	6	11	25	33	32	3	4	4	8	5	3	4	4	10	5
9	6	10	16	37	40	4	4	4	4	4	3	4	4	4	4
10	5	8	15	31	28	4	4	4	5	5	3	4	4	4	5
11	5	11	17	18	36	4	4	4	4	4	3	4	4	4	4
12	6	7	11	15	22	4	4	4	4	4	3	4	4	4	4
13	5	8	9	14	58	3	4	4	4	4	3	4	4	4	4
14	5	6	10	13	17	3	4	4	4	4	3	4	4	4	4
15	5	7	9	10	13	3	4	4	4	4	3	4	4	4	4
16	5	9	7	10	12	3	4	4	4	4	3	4	4	4	4
17	5	6	7	9	17	3	4	4	4	4	3	4	4	4	4
18	5	6	7	9	13	3	4	4	4	4	3	4	4	4	4
19	5	6	7	8	11	4	4	4	4	4	3	4	4	4	4
20	5	6	8	9	10	3	4	4	4	4	3	4	4	4	4
21	6	6	7	8	10	3	4	4	4	4	3	3	4	4	4
22	5	5	8	9	11	3	4	4	4	4	3	3	4	4	4
23	4	5	7	8	10	3	4	4	4	4	3	4	4	4	4
24	5	6	7	8	8	3	4	4	4	4	3	3	4	4	4
25	4	6	6	8	9	3	4	4	4	4	3	3	4	4	4
26	4	5	6	7	9	3	4	4	4	4	3	3	4	4	4
27	5	5	7	8	9	3	4	4	4	4	3	3	4	4	4
28	4	5	6	7	8	3	4	4	4	4	3	3	4	4	4
29	4	6	6	7	8	3	4	4	4	4	3	3	4	4	4
30	5	6	6	7	8	3	4	4	4	4	3	3	4	4	4
31	5	5	6	7	8	3	4	4	4	4	3	3	4	4	4
32	5	5	6	7	8	3	4	4	4	4	3	3	4	4	4
33	4	5	6	7	7	3	4	4	4	4	3	3	4	4	4
34	4	5	6	7	7	3	4	4	4	4	3	3	4	4	4
35	4	6	6	7	8	3	4	4	4	4	3	3	4	4	4
36	4	5	6	7	8	3	4	4	4	4	3	3	3	4	4
37	4	6	6	6	7	3	4	4	4	4	3	3	4	4	4
38	5	5	6	6	7	3	3	4	4	4	3	3	4	4	4
39	4	6	6	7	7	3	4	4	4	4	3	3	3	4	4
40	4	5	6	6	7	3	3	4	4	4	3	3	3	4	4
41	4	5	5	6	7	3	3	4	4	4	3	3	4	4	4
42	4	5	6	6	7	3	4	4	4	4	3	3	3	4	4
43	4	5	6	6	6	3	4	4	4	4	3	3	4	4	4
44	4	5	5	6	6	3	3	4	4	4	3	3	3	4	4
45	4	5	6	6	6	3	3	4	4	4	3	3	3	4	4
46	4	5	5	6	6	3	4	4	4	4	3	3	3	4	4
47	4	5	5	6	6	3	4	4	4	4	3	3	3	4	4
48	4	5	5	7	6	3	3	4	4	4	3	3	3	4	4
49	4	5	6	6	6	3	4	4	4	4	3	3	3	4	4
50	4	5	5	6	6	3	4	4	4	4	3	3	3	3	4

Table 2.2: Number of iterations for normally distributed data with large variances.

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	33	4	4	4	4	5	4	4	4	4	5	4	4	4	4
3	28	42	5	4	4	5	8	5	4	4	4	8	4	4	4
4	9	34	18	–	5	4	5	5	5	5	4	6	5	5	4
5	9	21	36	68	–	5	5	7	6	5	4	5	8	6	5
6	8	17	33	66	76	4	5	5	5	6	5	5	5	5	6
7	7	17	117	31	22	4	5	6	6	6	4	5	5	7	6
8	6	22	15	47	51	4	9	5	5	6	4	9	5	5	6
9	8	22	14	25	26	4	15	5	6	6	4	13	5	6	7
10	7	8	15	21	55	4	4	5	5	6	4	4	5	5	6
11	6	8	22	18	46	4	4	5	5	5	4	4	6	5	6
12	6	8	12	18	30	4	5	5	5	5	4	4	5	5	5
13	6	8	10	14	20	4	4	5	5	5	4	4	4	5	5
14	6	8	10	13	17	4	4	4	5	5	4	4	4	5	5
15	5	7	11	14	14	4	4	5	5	5	4	4	5	5	5
16	6	7	13	11	14	4	4	5	5	5	4	4	4	5	5
17	6	7	9	10	13	4	4	5	5	5	4	4	4	6	5
18	6	7	9	11	17	4	4	5	5	5	4	4	4	5	5
19	5	7	9	11	11	4	4	5	6	5	4	4	4	5	5
20	6	6	8	10	11	4	4	5	5	5	4	4	4	5	5
21	5	7	8	9	11	4	4	4	5	5	4	4	4	4	5
22	5	6	8	8	11	4	4	4	5	5	4	4	4	4	5
23	6	6	7	9	11	4	4	4	5	5	4	4	4	5	5
24	5	6	7	8	10	4	4	4	5	5	4	4	4	4	5
25	5	6	7	9	9	4	4	4	5	5	4	4	4	4	5
26	5	6	7	8	9	4	4	4	5	5	4	4	4	4	5
27	5	6	7	8	10	4	4	4	5	5	4	4	4	4	4
28	6	6	7	8	9	4	4	4	4	5	4	4	4	4	5
29	6	6	7	9	9	4	4	4	4	5	4	4	4	4	4
30	5	6	7	8	9	4	4	4	4	5	4	4	4	4	4
31	5	5	6	7	8	4	4	4	4	5	4	4	4	4	5
32	5	6	6	13	8	4	4	4	4	5	4	4	4	4	4
33	5	6	6	7	8	4	4	4	4	5	4	4	4	4	4
34	5	6	6	7	7	4	4	4	4	4	4	4	4	4	4
35	5	6	6	7	8	4	4	4	4	5	4	4	4	4	4
36	5	6	6	7	7	4	4	4	4	4	4	4	4	4	4
37	5	6	6	7	7	4	4	4	4	4	4	4	4	4	4
38	5	6	6	7	7	4	4	4	4	4	3	4	4	4	4
39	5	6	6	7	8	4	4	4	4	5	4	4	4	4	4
40	5	5	6	7	7	4	4	4	4	4	4	4	4	4	4
41	5	5	6	7	7	4	4	4	4	5	4	4	4	4	4
42	5	5	6	7	7	4	4	4	4	4	4	4	4	4	4
43	5	5	6	7	7	4	4	4	4	4	4	4	4	4	4
44	4	5	6	6	7	4	4	4	4	4	3	4	4	4	4
45	5	5	7	6	7	4	4	4	4	4	4	4	4	4	4
46	5	5	6	6	7	4	4	4	4	4	3	4	4	4	4
47	5	5	6	6	11	4	4	4	4	4	4	4	4	4	4
48	5	5	6	6	7	4	4	4	4	4	3	4	4	4	4
49	4	5	6	6	7	4	4	4	4	4	3	4	4	4	4
50	4	5	6	6	7	4	4	4	4	4	3	4	4	4	4

Table 2.3: Number of iterations for Student t distributed data with $\nu = 3$ and small variances.

Notes: Dashes denote cases where no convergence is reached within 1000 iterations.

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	62	4	4	4	4	4	4	4	4	4	4	4	4	4	4
3	23	20	4	4	4	5	5	5	4	4	4	5	5	4	4
4	27	32	20	699	5	17	29	5	5	5	12	15	5	5	4
5	8	30	36	68	–	5	18	27	6	5	4	21	26	6	5
6	7	35	29	25	33	4	5	5	5	7	4	5	5	5	7
7	8	13	57	20	33	4	5	5	5	7	4	4	6	5	6
8	7	12	48	76	30	4	5	5	6	7	4	4	5	5	8
9	7	9	14	68	31	4	5	5	5	6	4	4	5	5	6
10	7	9	14	17	29	4	4	5	5	6	4	4	5	5	6
11	6	12	14	27	35	4	12	5	5	6	4	14	5	5	5
12	6	8	12	19	22	4	4	5	5	6	4	4	4	5	6
13	6	8	13	13	18	4	4	5	5	5	4	4	5	5	5
14	6	8	9	16	19	4	4	5	5	5	4	4	4	5	5
15	5	8	9	13	14	4	4	5	5	5	4	4	4	5	5
16	6	7	9	10	15	4	4	5	5	5	4	4	5	5	5
17	6	7	9	10	13	4	4	4	5	5	4	4	4	5	5
18	5	6	9	10	12	4	4	5	5	5	4	4	4	5	5
19	5	7	8	18	13	4	4	4	5	5	4	4	4	5	5
20	5	6	8	10	11	4	4	4	5	5	4	4	4	5	5
21	5	11	8	9	11	4	4	5	5	5	4	4	4	4	5
22	5	6	7	13	11	4	4	4	5	5	4	4	4	5	5
23	5	6	8	9	10	4	4	4	5	5	4	4	4	4	5
24	5	6	7	8	10	4	4	4	5	5	4	4	4	4	5
25	5	7	8	8	10	4	4	4	5	5	4	4	4	4	4
26	5	6	7	8	9	4	4	4	4	5	4	4	4	4	4
27	5	6	7	8	9	4	4	4	4	5	4	4	4	4	5
28	5	6	7	9	9	4	4	4	4	5	4	4	4	4	4
29	5	6	7	7	9	4	4	4	4	5	4	4	4	4	4
30	5	6	7	8	9	4	4	4	4	5	4	4	4	4	5
31	5	6	7	7	8	4	4	4	4	5	4	4	4	4	4
32	5	6	6	7	8	4	4	4	4	5	4	4	4	4	5
33	5	6	7	8	9	4	4	4	4	5	3	4	4	4	4
34	5	6	6	7	8	4	4	4	4	4	4	4	4	4	4
35	5	6	6	7	8	4	4	4	4	5	4	4	4	4	4
36	5	8	6	7	8	4	4	4	4	5	4	4	4	4	4
37	8	6	11	7	7	4	4	4	4	4	4	4	4	4	4
38	5	5	6	7	7	4	4	4	4	4	4	4	4	4	4
39	5	6	6	7	7	4	4	4	4	4	4	4	4	4	4
40	5	5	6	6	7	4	4	4	4	4	3	4	4	4	4
41	5	6	6	7	7	4	4	4	4	4	4	4	4	4	4
42	5	5	6	6	7	4	4	4	4	4	3	4	4	4	4
43	5	5	6	7	7	4	4	4	4	4	3	4	4	4	4
44	4	5	6	7	7	4	4	4	4	4	3	4	4	4	4
45	5	5	6	6	7	4	4	4	4	4	3	4	4	4	4
46	5	5	6	7	7	4	4	4	4	4	3	4	4	4	4
47	5	5	6	7	7	4	4	4	4	4	3	4	4	4	4
48	5	5	6	6	7	4	4	4	4	4	4	4	4	4	4
49	4	5	6	6	7	4	4	4	4	4	3	4	4	4	4
50	4	5	6	6	7	4	4	4	4	4	3	4	4	4	4

Table 2.4: Number of iterations for Student t distributed data with $\nu = 3$ and large variances.

Notes: Dashes denote cases where no convergence is reached within 1000 iterations.

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	40	3	4	4	3	4	3	4	4	4	4	4	4	4	3
3	49	18	4	4	4	4	4	4	4	4	4	4	4	4	4
4	8	20	15	4	4	4	4	4	4	4	4	4	4	4	4
5	7	28	47	57	944	4	4	19	5	4	4	4	20	5	4
6	7	13	30	34	42	4	4	4	5	5	4	4	4	5	5
7	6	10	24	31	43	4	4	4	4	39	3	4	4	4	43
8	6	9	14	41	32	4	4	4	5	27	3	4	4	5	23
9	6	9	12	24	24	4	4	4	9	5	4	4	4	13	5
10	6	8	11	16	26	4	4	4	4	5	3	4	4	4	5
11	5	9	12	13	45	4	4	4	4	4	3	4	4	4	4
12	6	7	8	13	16	4	4	4	4	4	3	4	4	4	4
13	5	7	13	10	15	4	4	4	4	4	3	4	4	4	4
14	5	7	8	11	13	3	4	4	4	4	3	4	4	4	4
15	5	6	9	10	15	4	4	4	4	4	3	4	4	4	4
16	5	7	8	9	14	3	4	4	4	4	3	4	4	4	4
17	5	6	8	10	14	4	4	4	4	4	3	4	4	4	4
18	5	6	10	11	20	3	4	4	4	4	3	4	4	4	4
19	5	7	7	8	9	4	4	4	4	4	3	4	4	4	4
20	5	6	7	8	12	3	4	4	4	4	3	4	4	4	4
21	5	6	8	8	11	4	4	4	4	4	3	4	4	4	4
22	5	6	7	8	9	3	4	4	4	4	3	4	4	4	4
23	5	6	7	8	11	3	4	4	4	4	3	4	4	4	4
24	5	6	7	8	9	3	4	4	4	4	3	4	4	4	4
25	4	6	7	8	9	3	4	4	4	4	3	4	4	4	4
26	5	6	6	7	8	4	4	4	4	4	3	4	4	4	4
27	5	5	6	7	8	3	4	4	4	4	3	4	4	4	4
28	4	5	6	8	8	3	4	4	4	4	3	3	4	4	4
29	5	5	6	8	8	4	4	4	4	4	3	4	4	4	4
30	5	5	6	7	8	4	4	4	4	4	3	4	4	4	4
31	4	5	6	7	8	3	4	4	4	4	3	3	4	4	4
32	5	5	6	7	7	3	4	4	4	4	3	3	4	4	4
33	4	5	6	7	8	3	4	4	4	4	3	3	4	4	4
34	5	5	6	6	7	3	4	4	4	4	3	3	4	4	4
35	5	5	6	7	7	3	4	4	4	4	3	3	4	4	4
36	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
37	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
38	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
39	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
40	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
41	4	5	5	6	7	3	4	4	4	4	3	3	4	4	4
42	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
43	4	5	5	6	7	3	4	4	4	4	3	3	4	4	4
44	4	5	6	6	6	3	4	4	4	4	3	3	4	4	4
45	4	5	5	6	6	3	4	4	4	4	3	3	4	4	4
46	4	5	5	6	7	3	4	4	4	4	3	3	4	4	4
47	4	5	5	6	6	3	3	4	4	4	3	3	4	4	4
48	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
49	4	5	5	6	6	3	4	4	4	4	3	3	4	4	4
50	4	5	5	6	7	3	3	4	4	4	3	3	4	4	4

Table 2.5: Number of iterations for Student t distributed data with $\nu = 10$ and small variances.

Notes: Dashes denote cases where no convergence is reached within 1000 iterations.

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	39	4	4	4	4	4	4	4	4	4	4	4	4	3	3
3	8	20	4	4	4	4	4	4	4	4	4	4	4	4	4
4	24	33	16	4	4	13	4	5	4	4	15	4	5	4	4
5	9	36	66	65	852	4	17	16	5	5	4	15	18	5	5
6	6	32	29	17	45	4	11	4	19	9	4	9	4	12	9
7	7	11	18	40	53	4	4	4	5	14	4	4	4	5	13
8	5	9	38	26	33	4	4	4	5	8	4	4	4	4	8
9	7	11	14	28	18	4	4	4	4	6	4	4	4	4	8
10	5	8	11	45	27	4	4	4	4	4	3	4	4	4	4
11	5	8	12	14	23	4	4	4	4	5	3	4	4	4	5
12	5	6	10	20	30	3	4	4	4	4	3	4	4	4	4
13	6	8	11	15	13	4	4	4	4	4	3	4	4	4	4
14	5	7	8	10	17	4	4	4	4	4	3	4	4	4	4
15	5	6	9	10	16	4	4	4	4	4	3	4	4	4	4
16	6	7	9	9	12	4	4	4	4	4	3	4	4	4	4
17	5	7	8	9	12	3	4	4	4	4	3	4	4	4	4
18	5	6	9	9	14	3	4	4	4	4	3	4	4	4	4
19	5	6	7	9	12	4	4	4	4	4	3	4	4	4	4
20	4	6	7	8	10	3	4	4	4	4	3	4	4	4	4
21	5	6	7	8	9	3	4	4	4	4	3	4	4	4	4
22	5	6	7	7	9	3	4	4	4	4	3	4	4	4	4
23	5	6	7	7	9	3	4	4	4	4	3	4	4	4	4
24	5	6	6	7	10	4	4	4	4	4	3	4	4	4	4
25	4	6	6	8	9	3	4	4	4	4	3	4	4	4	4
26	5	5	7	8	8	3	4	4	4	4	3	4	4	4	4
27	4	6	6	7	8	3	4	4	4	4	3	3	4	4	4
28	5	6	6	8	8	4	4	4	4	4	3	3	4	4	4
29	5	5	6	7	8	3	4	4	4	4	3	4	4	4	4
30	5	5	6	7	7	3	4	4	4	4	3	3	4	4	4
31	4	5	7	7	8	3	4	4	4	4	3	3	4	4	4
32	5	5	6	7	7	3	4	4	4	4	3	3	4	4	4
33	4	5	6	7	8	3	4	4	4	4	3	3	4	4	4
34	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
35	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
36	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
37	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
38	5	5	7	6	7	3	4	4	4	4	3	3	4	4	4
39	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
40	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
41	4	5	5	6	7	3	4	4	4	4	3	3	4	4	4
42	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
43	4	5	5	6	7	3	4	4	4	4	3	3	4	4	4
44	4	5	5	6	7	3	3	4	4	4	3	3	4	4	4
45	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
46	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
47	4	5	5	6	6	3	4	4	4	4	3	3	4	4	4
48	4	5	5	6	6	3	4	4	4	4	3	3	4	4	4
49	4	5	5	6	6	3	4	4	4	4	3	3	4	4	4
50	4	5	6	6	7	3	4	4	4	4	3	3	3	4	4

Table 2.6: Number of iterations for Student t distributed data with $\nu = 10$ and large variances.

Notes: Dashes denote cases where no convergence is reached within 1000 iterations.

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	43	4	3	4	4	3	4	3	4	4	3	4	3	4	4
3	14	18	4	4	4	4	4	4	4	4	4	4	4	4	4
4	10	70	50	4	4	4	4	4	4	4	4	4	4	4	4
5	9	20	23	27	–	4	4	4	4	4	4	4	4	4	4
6	7	16	47	46	47	4	4	5	4	8	4	4	4	4	8
7	6	16	16	33	31	4	4	4	4	15	3	4	4	4	13
8	7	14	12	26	32	4	4	4	4	5	4	4	4	4	5
9	5	12	12	28	21	3	4	4	4	5	3	4	4	4	5
10	6	7	13	31	18	4	4	4	4	4	3	4	4	4	4
11	5	7	11	16	25	4	4	4	4	4	3	4	4	4	5
12	5	7	10	18	16	3	4	4	4	4	3	4	4	4	4
13	5	7	9	15	18	3	4	4	4	4	3	4	4	4	4
14	5	6	11	14	14	4	4	4	4	4	3	4	4	4	4
15	5	7	8	9	12	3	4	4	4	4	3	4	4	4	4
16	5	6	10	11	11	3	4	4	4	4	3	4	4	4	4
17	5	6	7	9	13	4	4	4	4	4	3	4	4	4	4
18	5	6	9	9	10	4	4	4	4	4	3	4	4	4	4
19	5	6	9	10	10	3	4	4	4	4	3	4	4	4	4
20	5	6	7	8	10	3	4	4	4	4	3	4	4	4	4
21	5	6	7	8	9	3	4	4	4	4	3	3	4	4	4
22	5	6	6	8	10	3	4	4	4	4	3	3	4	4	4
23	5	6	7	7	8	3	4	4	4	4	3	3	4	4	4
24	5	6	7	7	9	3	4	4	4	4	3	3	4	4	4
25	5	6	6	7	9	3	4	4	4	4	3	3	4	4	4
26	4	6	7	7	8	3	4	4	4	4	3	3	4	4	4
27	5	6	6	7	8	3	4	4	4	4	3	3	4	4	4
28	5	5	6	7	8	3	4	4	4	4	3	3	4	4	4
29	5	6	6	7	8	3	4	4	4	4	3	3	4	4	4
30	4	5	6	7	8	3	4	4	4	4	3	3	4	4	4
31	4	5	6	6	8	3	3	4	4	4	3	3	4	4	4
32	4	5	6	7	8	3	4	4	4	4	3	3	4	4	4
33	4	5	6	6	7	3	3	4	4	4	3	3	4	4	4
34	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
35	5	5	6	6	7	3	4	4	4	4	3	3	4	4	4
36	5	5	6	6	7	3	4	4	4	4	3	3	4	4	4
37	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
38	4	5	6	7	7	3	4	4	4	4	3	3	3	4	4
39	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
40	4	5	5	6	7	3	4	4	4	4	3	3	4	4	4
41	4	5	5	6	7	3	4	4	4	4	3	3	3	4	4
42	4	5	5	6	7	3	4	4	4	4	3	3	4	4	4
43	4	5	6	6	6	3	3	4	4	4	3	3	3	4	4
44	4	5	5	6	7	3	4	4	4	4	3	3	3	4	4
45	4	5	5	6	6	3	3	4	4	4	3	3	3	4	4
46	4	5	6	6	7	3	3	4	4	4	3	3	3	4	4
47	4	5	5	6	7	3	4	4	4	4	3	3	3	4	4
48	4	5	5	6	6	3	4	4	4	4	3	3	3	4	4
49	4	5	5	6	6	3	3	4	4	4	3	3	3	4	4
50	4	5	5	6	6	3	3	4	4	4	3	3	3	4	4

Table 2.7: Number of iterations for Student t distributed data with $\nu = 25$ and small variances.

Notes: Dashes denote cases where no convergence is reached within 1000 iterations.

K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	52	4	4	4	4	4	4	4	4	4	4	4	4	3	4
3	21	14	4	4	4	4	4	4	4	4	4	4	4	4	3
4	9	36	14	4	4	4	20	4	4	4	4	23	5	4	4
5	10	36	56	24	–	4	17	4	5	5	4	10	4	5	4
6	6	12	31	35	89	4	4	4	5	6	4	4	4	5	6
7	7	10	25	22	36	4	4	4	5	4	4	4	4	5	4
8	5	10	25	37	49	4	4	4	4	20	3	4	4	4	21
9	6	8	15	24	40	4	4	4	4	4	3	4	4	4	4
10	6	9	12	22	29	4	4	4	4	5	3	4	4	4	5
11	6	8	10	17	70	4	4	4	4	5	3	4	4	4	5
12	6	7	8	19	20	4	4	4	4	4	3	4	4	4	4
13	6	8	11	12	15	4	4	4	4	4	3	4	4	4	4
14	5	6	9	12	17	4	4	4	4	4	3	4	4	4	4
15	6	6	8	10	15	4	4	4	4	4	3	4	4	4	4
16	5	6	8	9	14	3	4	4	4	4	3	4	4	4	4
17	5	6	8	9	12	4	4	4	4	4	3	4	4	4	4
18	5	6	7	8	10	3	4	4	4	4	3	4	4	4	4
19	5	6	7	9	11	3	4	4	4	4	3	4	4	4	4
20	5	6	7	9	11	3	4	4	4	4	3	4	4	4	4
21	5	6	7	8	10	4	4	4	4	4	3	3	4	4	4
22	4	5	7	8	10	3	4	4	4	4	3	3	4	4	4
23	5	6	7	7	11	4	4	4	4	4	3	3	4	4	4
24	5	6	6	8	9	3	4	4	4	4	3	4	4	4	4
25	6	5	7	7	9	4	4	4	4	4	3	3	4	4	4
26	4	6	6	7	8	3	4	4	4	4	3	3	4	4	4
27	4	5	6	7	9	3	4	4	4	4	3	3	4	4	4
28	4	6	7	7	8	3	4	4	4	4	3	3	4	4	4
29	5	6	6	7	7	3	4	4	4	4	3	3	4	4	4
30	5	5	6	7	8	4	4	4	4	4	3	3	4	4	4
31	4	5	6	7	8	3	4	4	4	4	3	3	4	4	4
32	4	5	6	7	7	3	4	4	4	4	3	3	4	4	4
33	4	5	6	7	7	3	4	4	4	4	3	3	4	4	4
34	4	5	6	7	7	3	4	4	4	4	3	3	4	4	4
35	4	5	6	6	7	3	4	4	4	4	3	3	4	4	4
36	4	5	6	7	7	3	4	4	4	4	3	3	4	4	4
37	5	5	6	7	7	3	4	4	4	4	3	3	4	4	4
38	4	6	6	6	7	3	4	4	4	4	3	3	4	4	4
39	4	6	6	6	7	3	4	4	4	4	3	3	3	4	4
40	4	5	6	6	6	3	4	4	4	4	3	3	3	4	4
41	4	5	5	6	7	3	4	4	4	4	3	3	3	4	4
42	4	5	5	6	7	3	4	4	4	4	3	3	3	4	4
43	4	5	5	6	7	3	4	4	4	4	3	3	3	4	4
44	4	5	5	7	6	3	4	4	4	4	3	3	3	4	4
45	5	5	6	6	7	3	3	4	4	4	3	3	3	4	4
46	4	5	6	6	6	3	4	4	4	4	3	3	3	4	4
47	4	5	5	6	6	3	4	4	4	4	3	3	3	4	4
48	4	5	6	6	7	3	3	4	4	4	3	3	3	4	4
49	4	5	5	6	6	3	4	4	4	4	3	3	3	4	4
50	4	5	6	6	6	3	4	4	4	4	3	3	3	4	4

Table 2.8: Number of iterations for Student t distributed data with $\nu = 25$ and large variances.

Notes: Dashes denote cases where no convergence is reached within 1000 iterations.

small variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0020	0.0026	0.0066	0.0048	0.0293	0.0023	0.0023	0.0042	0.0065	0.0293	0.0018	0.0027	0.0072	0.0048	0.0292
10	0.0028	0.0027	0.0029	0.0026	0.0033	0.0028	0.0025	0.0030	0.0030	0.0039	0.0028	0.0027	0.0028	0.0027	0.0033
20	0.0028	0.0028	0.0023	0.0025	0.0026	0.0028	0.0028	0.0023	0.0025	0.0027	0.0028	0.0028	0.0023	0.0025	0.0026
30	0.0027	0.0024	0.0027	0.0023	0.0026	0.0027	0.0023	0.0026	0.0023	0.0026	0.0027	0.0024	0.0026	0.0022	0.0026
40	0.0026	0.0024	0.0023	0.0026	0.0026	0.0026	0.0024	0.0023	0.0026	0.0027	0.0026	0.0024	0.0023	0.0026	0.0026
50	0.0025	0.0023	0.0025	0.0022	0.0024	0.0025	0.0023	0.0025	0.0022	0.0024	0.0024	0.0023	0.0026	0.0022	0.0024
mean	0.0029	0.0035	0.0031	0.0029	0.0035	0.0030	0.0035	0.0031	0.0030	0.0035	0.0029	0.0035	0.0031	0.0029	0.0035
small variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0020	0.0026	0.0062	0.0048	0.0292	0.0025	0.0023	0.0048	0.0049	0.0291	0.0025	0.0023	0.0048	0.0049	0.0291
10	0.0028	0.0027	0.0029	0.0026	0.0033	0.0028	0.0024	0.0031	0.0027	0.0034	0.0028	0.0024	0.0031	0.0027	0.0034
20	0.0028	0.0028	0.0023	0.0025	0.0026	0.0028	0.0028	0.0021	0.0026	0.0028	0.0028	0.0028	0.0021	0.0026	0.0028
30	0.0027	0.0024	0.0027	0.0023	0.0026	0.0027	0.0024	0.0027	0.0024	0.0025	0.0027	0.0024	0.0027	0.0024	0.0025
40	0.0026	0.0024	0.0023	0.0026	0.0026	0.0026	0.0024	0.0023	0.0027	0.0028	0.0026	0.0024	0.0023	0.0027	0.0028
50	0.0025	0.0023	0.0025	0.0022	0.0024	0.0025	0.0023	0.0026	0.0022	0.0024	0.0025	0.0023	0.0026	0.0022	0.0024
mean	0.0028	0.0035	0.0031	0.0029	0.0035	0.0029	0.0035	0.0031	0.0030	0.0035	0.0029	0.0035	0.0031	0.0030	0.0035
large variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0087	0.0266	0.0276	0.0374	0.0464	0.0098	0.0239	0.0294	0.0376	0.0458	0.0090	0.0270	0.0273	0.0373	0.0464
10	0.0078	0.0116	0.0162	0.0204	0.0265	0.0079	0.0101	0.0155	0.0282	0.0251	0.0078	0.0117	0.0166	0.0205	0.0266
20	0.0106	0.0104	0.0117	0.0113	0.0143	0.0107	0.0101	0.0117	0.0112	0.0142	0.0106	0.0107	0.0117	0.0113	0.0143
30	0.0077	0.0086	0.0087	0.0094	0.0109	0.0080	0.0083	0.0089	0.0094	0.0107	0.0078	0.0088	0.0087	0.0093	0.0110
40	0.0071	0.0080	0.0088	0.0086	0.0096	0.0069	0.0079	0.0086	0.0085	0.0093	0.0073	0.0079	0.0088	0.0086	0.0096
50	0.0071	0.0082	0.0084	0.0087	0.0090	0.0072	0.0083	0.0085	0.0084	0.0088	0.0072	0.0081	0.0084	0.0087	0.0089
mean	0.0153	0.0190	0.0197	0.0171	0.0266	0.0149	0.0179	0.0198	0.0168	0.0263	0.0154	0.0192	0.0198	0.0171	0.0265
large variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0087	0.0265	0.0276	0.0374	0.0463	0.0099	0.0245	0.0259	0.0326	0.0441	0.0099	0.0245	0.0259	0.0326	0.0441
10	0.0078	0.0116	0.0161	0.0203	0.0265	0.0084	0.0109	0.0163	0.0205	0.0260	0.0084	0.0109	0.0163	0.0205	0.0260
20	0.0106	0.0104	0.0117	0.0113	0.0143	0.0108	0.0102	0.0122	0.0117	0.0145	0.0108	0.0102	0.0122	0.0117	0.0145
30	0.0077	0.0086	0.0087	0.0094	0.0109	0.0083	0.0084	0.0089	0.0102	0.0116	0.0083	0.0084	0.0089	0.0102	0.0116
40	0.0071	0.0079	0.0088	0.0086	0.0096	0.0069	0.0082	0.0089	0.0085	0.0100	0.0069	0.0082	0.0089	0.0085	0.0100
50	0.0071	0.0082	0.0084	0.0087	0.0090	0.0075	0.0085	0.0088	0.0084	0.0092	0.0075	0.0085	0.0088	0.0084	0.0092
mean	0.0152	0.0178	0.0197	0.0171	0.0228	0.0153	0.0177	0.0199	0.0169	0.0258	0.0153	0.0177	0.0199	0.0169	0.0258

Table 2.9: Average of the root mean-squared error (RMSE) over all nK entries of $\tilde{M}$ for normally distributed data.

small variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0398	0.0617	0.0860	0.1300	0.1477	0.0390	0.0513	0.0969	0.1059	0.1376	0.0398	0.0619	0.0857	0.1302	0.1474
10	0.0198	0.0346	0.0434	0.0622	0.0786	0.0189	0.0352	0.0425	0.0609	0.0652	0.0207	0.0341	0.0433	0.0624	0.0785
20	0.0159	0.0171	0.0250	0.0303	0.0364	0.0159	0.0168	0.0248	0.0301	0.0361	0.0156	0.0172	0.0247	0.0305	0.0361
30	0.0133	0.0148	0.0180	0.0205	0.0267	0.0136	0.0149	0.0180	0.0204	0.0265	0.0137	0.0149	0.0183	0.0205	0.0268
40	0.0155	0.0150	0.0173	0.0183	0.0203	0.0152	0.0148	0.0173	0.0185	0.0201	0.0155	0.0151	0.0169	0.0182	0.0203
50	0.0134	0.0143	0.0150	0.0163	0.0195	0.0133	0.0144	0.0149	0.0165	0.0194	0.0132	0.0145	0.0151	0.0160	0.0193
mean	0.0214	0.0425	0.0383	0.0473	0.0530	0.0221	0.0421	0.0395	0.0460	0.0521	0.0215	0.0427	0.0383	0.0473	0.0530
small variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0398	0.0617	0.0856	0.1294	0.1471	0.0432	0.0561	0.0831	0.1064	0.1373	0.0432	0.0561	0.0831	0.1064	0.1373
10	0.0197	0.0345	0.0433	0.0621	0.0784	0.0189	0.0355	0.0553	0.0568	0.0839	0.0189	0.0355	0.0553	0.0568	0.0839
20	0.0159	0.0170	0.0249	0.0303	0.0363	0.0168	0.0181	0.0261	0.0322	0.0416	0.0168	0.0181	0.0261	0.0322	0.0416
30	0.0133	0.0148	0.0179	0.0205	0.0267	0.0153	0.0149	0.0202	0.0235	0.0289	0.0153	0.0149	0.0202	0.0235	0.0289
40	0.0155	0.0150	0.0172	0.0183	0.0203	0.0152	0.0157	0.0183	0.0202	0.0229	0.0152	0.0157	0.0183	0.0202	0.0229
50	0.0133	0.0143	0.0150	0.0162	0.0194	0.0129	0.0147	0.0159	0.0193	0.0225	0.0129	0.0147	0.0159	0.0193	0.0225
mean	0.0214	0.0413	0.0381	0.0471	0.0526	0.0220	0.0441	0.0402	0.0492	0.0557	0.0220	0.0441	0.0402	0.0492	0.0557
large variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0393	0.1338	0.1818	0.1170	0.1615	0.0391	0.1537	0.1611	0.0924	0.1548	0.0394	0.1262	0.1849	0.1171	0.1614
10	0.0215	0.0251	0.0498	0.0598	0.0683	0.0212	0.0261	0.0493	0.0592	0.0666	0.0219	0.0262	0.0498	0.0599	0.0681
20	0.0156	0.0194	0.0245	0.0348	0.0377	0.0154	0.0189	0.0248	0.0344	0.0370	0.0155	0.0198	0.0245	0.0347	0.0377
30	0.0126	0.0167	0.0184	0.0219	0.0261	0.0127	0.0164	0.0181	0.0218	0.0257	0.0128	0.0161	0.0187	0.0221	0.0262
40	0.0151	0.0142	0.0180	0.0187	0.0205	0.0152	0.0142	0.0176	0.0187	0.0208	0.0150	0.0143	0.0178	0.0189	0.0204
50	0.0108	0.0144	0.0159	0.0164	0.0196	0.0107	0.0143	0.0159	0.0164	0.0197	0.0109	0.0145	0.0159	0.0163	0.0197
mean	0.0321	0.0396	0.0464	0.0480	0.0540	0.0330	0.0402	0.0461	0.0462	0.0524	0.0324	0.0395	0.0465	0.0480	0.0540
large variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0393	0.1203	0.1722	0.1170	0.1598	0.0426	0.1307	0.1878	0.1068	0.1436	0.0426	0.1307	0.1878	0.1068	0.1436
10	0.0215	0.0250	0.0498	0.0598	0.0682	0.0223	0.0320	0.0545	0.0677	0.0781	0.0223	0.0320	0.0545	0.0677	0.0781
20	0.0156	0.0194	0.0244	0.0348	0.0376	0.0158	0.0188	0.0298	0.0350	0.0411	0.0158	0.0188	0.0298	0.0350	0.0411
30	0.0126	0.0166	0.0184	0.0219	0.0260	0.0125	0.0166	0.0182	0.0244	0.0312	0.0125	0.0166	0.0182	0.0244	0.0312
40	0.0151	0.0142	0.0180	0.0186	0.0204	0.0157	0.0153	0.0204	0.0213	0.0220	0.0157	0.0153	0.0204	0.0213	0.0220
50	0.0108	0.0144	0.0158	0.0163	0.0196	0.0115	0.0149	0.0168	0.0170	0.0221	0.0115	0.0149	0.0168	0.0170	0.0221
mean	0.0303	0.0388	0.0458	0.0477	0.0536	0.0315	0.0408	0.0486	0.0499	0.0564	0.0315	0.0408	0.0486	0.0499	0.0564

Table 2.10: Average of the root mean-squared error (RMSE) over all nK entries of $\tilde{M}$ for Student t distributed data with $\nu = 3$.

small variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0112	0.0239	0.0984	0.0545	0.0695	0.0114	0.0235	0.1126	0.0581	0.0698	0.0119	0.0239	0.0988	0.0546	0.0697
10	0.0116	0.0148	0.0202	0.0243	0.0314	0.0117	0.0148	0.0195	0.0236	0.0316	0.0114	0.0152	0.0200	0.0246	0.0315
20	0.0082	0.0107	0.0118	0.0142	0.0161	0.0083	0.0108	0.0118	0.0141	0.0161	0.0083	0.0110	0.0117	0.0140	0.0161
30	0.0094	0.0092	0.0108	0.0112	0.0125	0.0095	0.0092	0.0106	0.0112	0.0127	0.0094	0.0092	0.0108	0.0112	0.0125
40	0.0090	0.0100	0.0100	0.0095	0.0117	0.0090	0.0100	0.0101	0.0096	0.0117	0.0090	0.0100	0.0099	0.0095	0.0119
50	0.0099	0.0089	0.0099	0.0095	0.0101	0.0099	0.0089	0.0099	0.0094	0.0101	0.0099	0.0087	0.0098	0.0095	0.0103
mean	0.0170	0.0141	0.0206	0.0203	0.0254	0.0172	0.0144	0.0214	0.0211	0.0268	0.0170	0.0141	0.0207	0.0203	0.0254
small variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0112	0.0239	0.0976	0.0544	0.0693	0.0122	0.0237	0.0982	0.0533	0.0686	0.0122	0.0237	0.0982	0.0533	0.0686
10	0.0116	0.0147	0.0201	0.0242	0.0314	0.0117	0.0158	0.0179	0.0240	0.0311	0.0117	0.0158	0.0179	0.0240	0.0311
20	0.0082	0.0107	0.0118	0.0142	0.0161	0.0084	0.0108	0.0121	0.0142	0.0175	0.0084	0.0108	0.0121	0.0142	0.0175
30	0.0094	0.0092	0.0108	0.0112	0.0124	0.0098	0.0100	0.0114	0.0126	0.0141	0.0098	0.0100	0.0114	0.0126	0.0141
40	0.0090	0.0100	0.0100	0.0095	0.0117	0.0091	0.0103	0.0106	0.0109	0.0127	0.0091	0.0103	0.0106	0.0109	0.0127
50	0.0099	0.0089	0.0099	0.0094	0.0101	0.0101	0.0091	0.0101	0.0102	0.0108	0.0101	0.0091	0.0101	0.0102	0.0108
mean	0.0169	0.0141	0.0205	0.0202	0.0253	0.0172	0.0149	0.0211	0.0208	0.0258	0.0172	0.0149	0.0211	0.0208	0.0258
large variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0182	0.0692	0.0760	0.0713	0.0985	0.0184	0.1197	0.0604	0.0707	0.0982	0.0180	0.0717	0.0775	0.0713	0.0983
10	0.0098	0.0134	0.0220	0.0229	0.0301	0.0100	0.0141	0.0219	0.0299	0.0299	0.0100	0.0136	0.0219	0.0221	0.0299
20	0.0090	0.0080	0.0122	0.0138	0.0175	0.0090	0.0083	0.0121	0.0139	0.0175	0.0089	0.0079	0.0121	0.0138	0.0175
30	0.0089	0.0090	0.0104	0.0104	0.0130	0.0091	0.0091	0.0105	0.0105	0.0129	0.0090	0.0091	0.0103	0.0104	0.0127
40	0.0102	0.0082	0.0087	0.0105	0.0112	0.0103	0.0083	0.0088	0.0103	0.0113	0.0101	0.0083	0.0088	0.0105	0.0113
50	0.0081	0.0090	0.0101	0.0112	0.0106	0.0081	0.0090	0.0101	0.0113	0.0106	0.0081	0.0091	0.0101	0.0113	0.0106
mean	0.0146	0.0162	0.0204	0.0231	0.0283	0.0154	0.0173	0.0205	0.0236	0.0285	0.0147	0.0163	0.0204	0.0231	0.0283
large variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0182	0.0621	0.0646	0.0696	0.0958	0.0179	0.0621	0.0629	0.0693	0.0953	0.0179	0.0621	0.0629	0.0693	0.0953
10	0.0098	0.0133	0.0220	0.0229	0.0301	0.0106	0.0161	0.0215	0.0249	0.0303	0.0106	0.0161	0.0215	0.0249	0.0303
20	0.0090	0.0079	0.0122	0.0137	0.0174	0.0093	0.0103	0.0126	0.0142	0.0180	0.0093	0.0103	0.0126	0.0142	0.0180
30	0.0089	0.0090	0.0104	0.0103	0.0130	0.0096	0.0098	0.0111	0.0111	0.0138	0.0096	0.0098	0.0111	0.0111	0.0138
40	0.0102	0.0082	0.0087	0.0105	0.0112	0.0104	0.0087	0.0094	0.0108	0.0116	0.0104	0.0087	0.0094	0.0108	0.0116
50	0.0081	0.0090	0.0101	0.0112	0.0106	0.0082	0.0093	0.0106	0.0119	0.0109	0.0082	0.0093	0.0106	0.0119	0.0109
mean	0.0142	0.0160	0.0201	0.0229	0.0279	0.0147	0.0166	0.0207	0.0236	0.0285	0.0147	0.0166	0.0207	0.0236	0.0285

Table 2.11: Average of the root mean-squared error (RMSE) over all nK entries of $\tilde{M}$ for Student t distributed data with $\nu = 10$.

small variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0060	0.0245	0.0324	0.0394	0.0614	0.0078	0.0260	0.0405	0.0560	0.0610	0.0052	0.0242	0.0325	0.0393	0.0614
10	0.0113	0.0133	0.0197	0.0205	0.0259	0.0118	0.0133	0.0195	0.0194	0.0317	0.0116	0.0135	0.0200	0.0207	0.0258
20	0.0068	0.0097	0.0098	0.0119	0.0158	0.0069	0.0097	0.0095	0.0117	0.0158	0.0069	0.0098	0.0098	0.0122	0.0158
30	0.0076	0.0081	0.0102	0.0095	0.0120	0.0077	0.0082	0.0103	0.0095	0.0121	0.0077	0.0082	0.0103	0.0097	0.0119
40	0.0086	0.0094	0.0097	0.0095	0.0104	0.0086	0.0094	0.0097	0.0095	0.0104	0.0086	0.0095	0.0097	0.0095	0.0103
50	0.0100	0.0086	0.0085	0.0097	0.0094	0.0100	0.0086	0.0085	0.0097	0.0093	0.0100	0.0086	0.0085	0.0097	0.0093
mean	0.0126	0.0155	0.0170	0.0185	0.0270	0.0126	0.0155	0.0173	0.0190	0.0276	0.0126	0.0156	0.0171	0.0185	0.0270
small variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0060	0.0244	0.0324	0.0394	0.0613	0.0130	0.0258	0.0350	0.0462	0.0613	0.0130	0.0258	0.0350	0.0462	0.0613
10	0.0113	0.0133	0.0197	0.0204	0.0259	0.0132	0.0135	0.0197	0.0213	0.0281	0.0132	0.0135	0.0197	0.0213	0.0281
20	0.0068	0.0097	0.0097	0.0119	0.0158	0.0073	0.0099	0.0109	0.0132	0.0153	0.0073	0.0099	0.0109	0.0132	0.0153
30	0.0076	0.0081	0.0102	0.0095	0.0119	0.0080	0.0085	0.0105	0.0107	0.0121	0.0080	0.0085	0.0105	0.0107	0.0121
40	0.0086	0.0094	0.0097	0.0095	0.0104	0.0085	0.0095	0.0102	0.0098	0.0106	0.0085	0.0095	0.0102	0.0098	0.0106
50	0.0100	0.0086	0.0085	0.0097	0.0093	0.0101	0.0089	0.0089	0.0104	0.0098	0.0101	0.0089	0.0089	0.0104	0.0098
mean	0.0126	0.0155	0.0170	0.0184	0.0266	0.0130	0.0161	0.0174	0.0192	0.0269	0.0130	0.0161	0.0174	0.0192	0.0269
large variance, $\tilde{M}$ unknown, iterate until convergence															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0122	0.0450	0.0311	0.1015	0.0671	0.0134	0.0674	0.0309	0.0952	0.0672	0.0113	0.0506	0.0310	0.1016	0.0673
10	0.0110	0.0149	0.0173	0.0242	0.0258	0.0110	0.0151	0.0174	0.0238	0.0239	0.0109	0.0147	0.0170	0.0242	0.0257
20	0.0075	0.0085	0.0120	0.0120	0.0154	0.0075	0.0086	0.0122	0.0120	0.0152	0.0077	0.0087	0.0119	0.0121	0.0156
30	0.0087	0.0083	0.0104	0.0098	0.0116	0.0087	0.0084	0.0105	0.0098	0.0116	0.0087	0.0083	0.0104	0.0098	0.0116
40	0.0088	0.0081	0.0080	0.0096	0.0113	0.0088	0.0081	0.0081	0.0097	0.0113	0.0088	0.0081	0.0081	0.0098	0.0113
50	0.0078	0.0079	0.0089	0.0087	0.0095	0.0079	0.0080	0.0089	0.0088	0.0096	0.0079	0.0079	0.0089	0.0087	0.0095
mean	0.0171	0.0143	0.0187	0.0219	0.0245	0.0171	0.0149	0.0193	0.0218	0.0249	0.0171	0.0144	0.0187	0.0219	0.0246
large variance, $\tilde{M}$ known, one iteration															
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n	equal weights					first scheme					second scheme				
5	0.0122	0.0393	0.0311	0.0977	0.0666	0.0167	0.0372	0.0330	0.0982	0.0679	0.0167	0.0372	0.0330	0.0982	0.0679
10	0.0109	0.0149	0.0172	0.0242	0.0258	0.0117	0.0153	0.0192	0.0233	0.0254	0.0117	0.0153	0.0192	0.0233	0.0254
20	0.0075	0.0085	0.0120	0.0119	0.0154	0.0080	0.0088	0.0142	0.0135	0.0168	0.0080	0.0088	0.0142	0.0135	0.0168
30	0.0087	0.0083	0.0104	0.0098	0.0116	0.0087	0.0086	0.0108	0.0108	0.0117	0.0087	0.0086	0.0108	0.0108	0.0117
40	0.0088	0.0081	0.0080	0.0096	0.0113	0.0088	0.0083	0.0086	0.0104	0.0114	0.0088	0.0083	0.0086	0.0104	0.0114
50	0.0078	0.0079	0.0089	0.0087	0.0095	0.0080	0.0081	0.0094	0.0091	0.0102	0.0080	0.0081	0.0094	0.0091	0.0102
mean	0.0170	0.0142	0.0186	0.0217	0.0243	0.0175	0.0147	0.0195	0.0224	0.0250	0.0175	0.0147	0.0195	0.0224	0.0250

Table 2.12: Average of the root mean-squared error (RMSE) over all nK entries of $\tilde{M}$ for Student t distributed data with $\nu = 25$.

								5 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	0.0379	0.0786	0.1227	0.1536	0.2042	0.0380	0.0772	0.1155	0.1476	0.1198	0.0353	0.0759	0.1177	0.1514	0.2032
	(0.0368)	(0.0877)	(0.1057)	(0.1473)	(0.1592)	(0.0371)	(0.0750)	(0.1534)	(0.1616)	(0.3625)	(0.0432)	(0.0815)	(0.1035)	(0.1533)	(0.1602)
30	0.0088	0.0230	0.0410	0.0524	0.0620	0.0084	0.0222	0.0400	0.0523	0.0609	0.0081	0.0230	0.0406	0.0519	0.0601
	(0.0137)	(0.0262)	(0.0367)	(0.0493)	(0.0601)	(0.0145)	(0.0248)	(0.0365)	(0.0464)	(0.0590)	(0.0140)	(0.0276)	(0.0378)	(0.0500)	(0.0584)
50	0.0087	0.0132	0.0212	0.0282	0.0379	0.0080	0.0130	0.0211	0.0271	0.0365	0.0082	0.0132	0.0212	0.0281	0.0373
	(0.0134)	(0.0181)	(0.0219)	(0.0296)	(0.0355)	(0.0129)	(0.0171)	(0.0228)	(0.0286)	(0.0374)	(0.0134)	(0.0184)	(0.0219)	(0.0288)	(0.0352)
								25 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	0.0136	0.0303	0.0474	0.0644	0.0851	0.0148	0.0295	0.0442	0.0599	0.0524	0.0139	0.0282	0.0448	0.0631	0.0843
	(0.0170)	(0.0399)	(0.0428)	(0.0646)	(0.0720)	(0.0163)	(0.0360)	(0.0628)	(0.0684)	(0.1522)	(0.0191)	(0.0363)	(0.0430)	(0.0681)	(0.0733)
30	0.0046	0.0097	0.0152	0.0211	0.0265	0.0041	0.0100	0.0155	0.0211	0.0262	0.0047	0.0100	0.0151	0.0204	0.0255
	(0.0069)	(0.0120)	(0.0164)	(0.0213)	(0.0257)	(0.0071)	(0.0116)	(0.0158)	(0.0211)	(0.0256)	(0.0071)	(0.0123)	(0.0165)	(0.0217)	(0.0253)
50	0.0029	0.0060	0.0096	0.0120	0.0153	0.0027	0.0059	0.0094	0.0115	0.0152	0.0034	0.0061	0.0096	0.0117	0.0152
	(0.0061)	(0.0079)	(0.0110)	(0.0135)	(0.0154)	(0.0057)	(0.0079)	(0.0107)	(0.0131)	(0.0158)	(0.0059)	(0.0082)	(0.0108)	(0.0134)	(0.0155)
								50 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	−0.0007	0.0009	−0.0014	−0.0001	−0.0017	0.0013	−0.0004	−0.0023	−0.0018	0.0034	−0.0009	0.0006	−0.0024	−0.0002	−0.0017
	(0.0093)	(0.0115)	(0.0192)	(0.0221)	(0.0291)	(0.0072)	(0.0108)	(0.0211)	(0.0208)	(0.0292)	(0.0117)	(0.0118)	(0.0195)	(0.0230)	(0.0297)
30	0.0012	0.0006	−0.0019	−0.0009	0.0001	0.0012	0.0003	−0.0020	−0.0001	−0.0001	0.0012	0.0008	−0.0019	−0.0011	−0.0003
	(0.0042)	(0.0061)	(0.0083)	(0.0101)	(0.0109)	(0.0043)	(0.0057)	(0.0081)	(0.0096)	(0.0106)	(0.0045)	(0.0062)	(0.0082)	(0.0101)	(0.0110)
50	0.0001	−0.0001	−0.0003	0.0003	0.0003	−0.0000	0.0002	−0.0004	−0.0000	0.0005	0.0002	0.0000	−0.0001	0.0003	0.0004
	(0.0046)	(0.0055)	(0.0061)	(0.0069)	(0.0076)	(0.0048)	(0.0050)	(0.0062)	(0.0066)	(0.0075)	(0.0047)	(0.0055)	(0.0063)	(0.0069)	(0.0076)
								75 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	−0.0162	−0.0298	−0.0500	−0.0649	−0.0866	−0.0135	−0.0304	−0.0487	−0.0640	−0.0445	−0.0153	−0.0303	−0.0499	−0.0638	−0.0862
	(0.0157)	(0.0355)	(0.0483)	(0.0586)	(0.0683)	(0.0145)	(0.0353)	(0.0694)	(0.0705)	(0.1370)	(0.0170)	(0.0370)	(0.0491)	(0.0599)	(0.0700)
30	−0.0032	−0.0096	−0.0176	−0.0219	−0.0262	−0.0032	−0.0099	−0.0165	−0.0211	−0.0261	−0.0034	−0.0093	−0.0170	−0.0213	−0.0260
	(0.0074)	(0.0114)	(0.0159)	(0.0212)	(0.0255)	(0.0077)	(0.0117)	(0.0159)	(0.0204)	(0.0252)	(0.0078)	(0.0114)	(0.0162)	(0.0215)	(0.0249)
50	−0.0029	−0.0060	−0.0091	−0.0110	−0.0152	−0.0032	−0.0058	−0.0092	−0.0110	−0.0140	−0.0032	−0.0058	−0.0087	−0.0109	−0.0152
	(0.0060)	(0.0075)	(0.0106)	(0.0131)	(0.0156)	(0.0060)	(0.0075)	(0.0105)	(0.0134)	(0.0157)	(0.0057)	(0.0076)	(0.0108)	(0.0131)	(0.0156)
								95 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	−0.0316	−0.0721	−0.1237	−0.1514	−0.2041	−0.0294	−0.0732	−0.1206	−0.1460	−0.1060	−0.0272	−0.0709	−0.1228	−0.1487	−0.2031
	(0.0391)	(0.0764)	(0.1063)	(0.1409)	(0.1564)	(0.0336)	(0.0711)	(0.1599)	(0.1625)	(0.3330)	(0.0424)	(0.0757)	(0.1081)	(0.1466)	(0.1613)
30	−0.0112	−0.0253	−0.0401	−0.0529	−0.0618	−0.0111	−0.0258	−0.0395	−0.0522	−0.0615	−0.0102	−0.0253	−0.0389	−0.0525	−0.0612
	(0.0145)	(0.0272)	(0.0363)	(0.0508)	(0.0575)	(0.0162)	(0.0264)	(0.0359)	(0.0469)	(0.0579)	(0.0153)	(0.0285)	(0.0363)	(0.0512)	(0.0560)
50	−0.0078	−0.0144	−0.0210	−0.0294	−0.0368	−0.0075	−0.0150	−0.0214	−0.0284	−0.0365	−0.0074	−0.0146	−0.0209	−0.0291	−0.0363
	(0.0126)	(0.0177)	(0.0239)	(0.0305)	(0.0355)	(0.0104)	(0.0163)	(0.0240)	(0.0312)	(0.0357)	(0.0127)	(0.0184)	(0.0242)	(0.0303)	(0.0351)

Table 2.13: Difference between simulated and postprocessed data quantiles for normally distributed data (standard errors in parentheses).

								5 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	0.0520	0.0507	0.1167	0.1593	0.2487	0.0485	0.0463	0.1220	0.1496	0.0194	0.0434	0.0470	0.1162	0.1589	0.2493
	(0.0375)	(0.0791)	(0.1089)	(0.1378)	(0.1828)	(0.0366)	(0.0717)	(0.1223)	(0.1399)	(0.5978)	(0.0404)	(0.0794)	(0.1093)	(0.1423)	(0.1920)
30	0.0114	0.0235	0.0337	0.0486	0.0621	0.0121	0.0239	0.0340	0.0495	0.0604	0.0125	0.0233	0.0329	0.0468	0.0625
	(0.0229)	(0.0339)	(0.0409)	(0.0572)	(0.0615)	(0.0240)	(0.0358)	(0.0458)	(0.0618)	(0.0667)	(0.0267)	(0.0323)	(0.0431)	(0.0592)	(0.0623)
50	0.0068	0.0136	0.0223	0.0313	0.0359	0.0074	0.0125	0.0217	0.0301	0.0355	0.0058	0.0115	0.0222	0.0302	0.0348
	(0.0201)	(0.0256)	(0.0300)	(0.0358)	(0.0396)	(0.0233)	(0.0256)	(0.0319)	(0.0395)	(0.0414)	(0.0207)	(0.0254)	(0.0312)	(0.0365)	(0.0422)
								25 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	−0.0119	−0.0177	−0.0168	−0.0042	0.0143	−0.0130	−0.0184	−0.0137	−0.0080	−0.0745	−0.0145	−0.0195	−0.0179	−0.0049	0.0151
	(0.0173)	(0.0281)	(0.0399)	(0.0468)	(0.0608)	(0.0173)	(0.0279)	(0.0407)	(0.0456)	(0.2390)	(0.0201)	(0.0290)	(0.0405)	(0.0471)	(0.0646)
30	−0.0076	−0.0107	−0.0175	−0.0189	−0.0207	−0.0084	−0.0110	−0.0177	−0.0196	−0.0215	−0.0081	−0.0110	−0.0174	−0.0196	−0.0212
	(0.0103)	(0.0108)	(0.0120)	(0.0137)	(0.0186)	(0.0107)	(0.0111)	(0.0140)	(0.0162)	(0.0207)	(0.0111)	(0.0115)	(0.0127)	(0.0158)	(0.0198)
50	−0.0044	−0.0097	−0.0105	−0.0151	−0.0178	−0.0045	−0.0098	−0.0110	−0.0159	−0.0181	−0.0041	−0.0100	−0.0111	−0.0158	−0.0183
	(0.0081)	(0.0093)	(0.0099)	(0.0116)	(0.0127)	(0.0090)	(0.0096)	(0.0104)	(0.0133)	(0.0131)	(0.0082)	(0.0093)	(0.0104)	(0.0126)	(0.0138)
								50 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	−0.0003	−0.0055	−0.0044	−0.0049	0.0022	−0.0008	−0.0050	−0.0022	−0.0062	−0.0042	0.0002	−0.0056	−0.0045	−0.0042	0.0041
	(0.0128)	(0.0132)	(0.0194)	(0.0269)	(0.0356)	(0.0126)	(0.0138)	(0.0185)	(0.0257)	(0.0652)	(0.0154)	(0.0144)	(0.0199)	(0.0265)	(0.0360)
30	0.0018	−0.0005	0.0007	−0.0009	−0.0006	0.0021	−0.0005	0.0006	−0.0007	−0.0006	0.0020	−0.0009	0.0009	−0.0011	−0.0003
	(0.0065)	(0.0088)	(0.0093)	(0.0108)	(0.0116)	(0.0063)	(0.0087)	(0.0098)	(0.0110)	(0.0121)	(0.0065)	(0.0084)	(0.0092)	(0.0107)	(0.0116)
50	0.0002	0.0004	−0.0009	0.0002	0.0006	0.0003	0.0004	−0.0011	−0.0000	0.0006	−0.0001	0.0001	−0.0009	0.0003	0.0008
	(0.0064)	(0.0071)	(0.0076)	(0.0093)	(0.0093)	(0.0062)	(0.0075)	(0.0075)	(0.0096)	(0.0088)	(0.0063)	(0.0069)	(0.0074)	(0.0090)	(0.0094)
								75 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	0.0138	0.0112	0.0033	−0.0039	−0.0103	0.0128	0.0130	0.0070	−0.0027	0.0622	0.0167	0.0131	0.0036	−0.0021	−0.0072
	(0.0176)	(0.0370)	(0.0331)	(0.0397)	(0.0641)	(0.0187)	(0.0343)	(0.0318)	(0.0509)	(0.2569)	(0.0228)	(0.0421)	(0.0353)	(0.0386)	(0.0627)
30	0.0094	0.0105	0.0150	0.0183	0.0227	0.0097	0.0104	0.0157	0.0184	0.0237	0.0098	0.0104	0.0155	0.0186	0.0230
	(0.0089)	(0.0117)	(0.0144)	(0.0152)	(0.0179)	(0.0093)	(0.0120)	(0.0152)	(0.0169)	(0.0200)	(0.0102)	(0.0117)	(0.0155)	(0.0176)	(0.0192)
50	0.0048	0.0078	0.0110	0.0143	0.0166	0.0056	0.0080	0.0113	0.0147	0.0169	0.0055	0.0082	0.0117	0.0149	0.0173
	(0.0077)	(0.0096)	(0.0103)	(0.0126)	(0.0127)	(0.0080)	(0.0099)	(0.0110)	(0.0136)	(0.0135)	(0.0082)	(0.0097)	(0.0109)	(0.0137)	(0.0138)
								95 % quantile							
K	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
n			equal weights					first scheme					second scheme		
10	−0.0275	−0.0806	−0.1298	−0.1847	−0.2387	−0.0315	−0.0803	−0.1206	−0.1779	−0.0525	−0.0262	−0.0796	−0.1300	−0.1813	−0.2312
	(0.0642)	(0.0872)	(0.1106)	(0.1332)	(0.1867)	(0.0681)	(0.0942)	(0.1003)	(0.1522)	(0.6822)	(0.0633)	(0.1026)	(0.1120)	(0.1237)	(0.1792)
30	−0.0104	−0.0277	−0.0347	−0.0512	−0.0563	−0.0108	−0.0263	−0.0343	−0.0513	−0.0559	−0.0078	−0.0262	−0.0345	−0.0497	−0.0557
	(0.0219)	(0.0346)	(0.0419)	(0.0504)	(0.0601)	(0.0241)	(0.0349)	(0.0424)	(0.0526)	(0.0675)	(0.0208)	(0.0324)	(0.0419)	(0.0522)	(0.0593)
50	−0.0037	−0.0176	−0.0249	−0.0302	−0.0393	−0.0034	−0.0183	−0.0248	−0.0294	−0.0392	−0.0023	−0.0166	−0.0244	−0.0285	−0.0383
	(0.0197)	(0.0244)	(0.0278)	(0.0375)	(0.0425)	(0.0213)	(0.0252)	(0.0288)	(0.0400)	(0.0448)	(0.0196)	(0.0245)	(0.0283)	(0.0392)	(0.0424)

Table 2.14: Difference between simulated and postprocessed data quantiles for Student t distributed data with $\nu = 3$ (standard errors in parentheses).

Figures

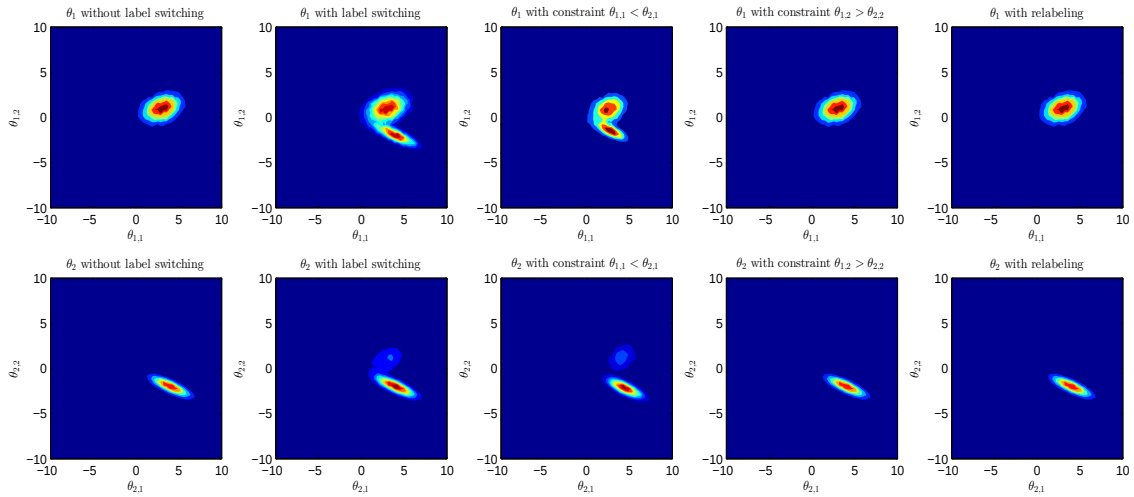

Figure 2.1: Example illustrating the effect of imposing ordering constraints.

Notes: A sample from the true, but unknown posterior distribution of the two bivariate parameters θ_1 and θ_2 is shown in the first column. The second column shows the available sample, which is subject to label switching. The third column shows the sample after imposing the ordering constraint $\theta_{1,1} < \theta_{2,1}$, the fourth column shows the sample after imposing the ordering constraint $\theta_{1,2} < \theta_{2,2}$, and the fifth column shows the sample after postprocessing it with the relabeling algorithm of Stephens (2000).

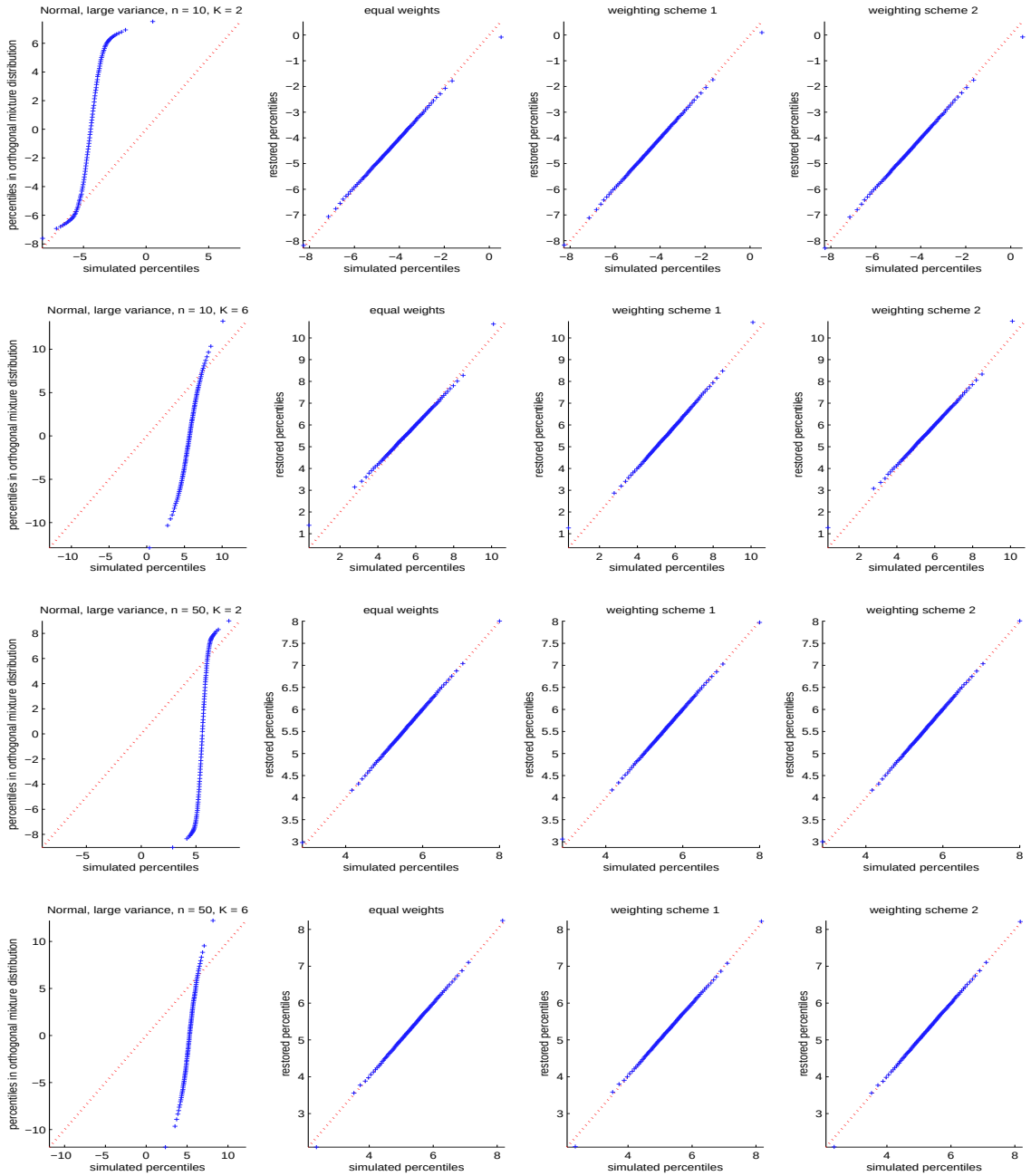

Figure 2.2: Orthogonally mixed and postprocessed data plotted against orthogonally invariant data for one randomly chosen $x_{j,k}$, normally distributed.

Notes: First row: $n = 10$ and $K = 2$, second row: $n = 10$ and $K = 6$, third row: $n = 50$ and $N = 2$, fourth row: $n = 50$ and $K = 6$.

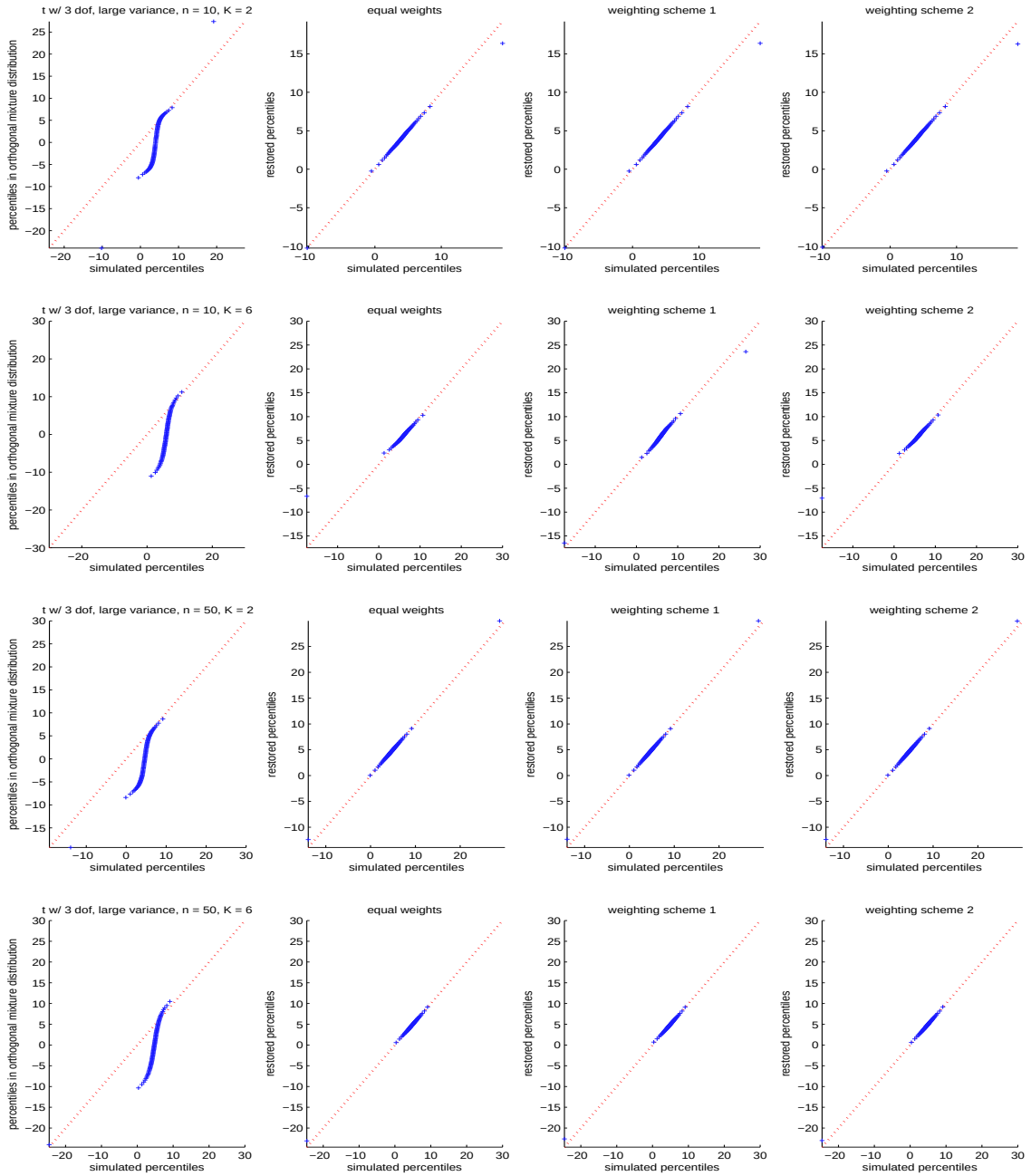

Figure 2.3: Orthogonally mixed and postprocessed data plotted against orthogonally invariant data for one randomly chosen $x_{j,k}$, Student t distributed with $\nu = 3$.

Notes: First row: $n = 10$ and $K = 2$, second row: $n = 10$ and $K = 6$, third row: $n = 50$ and $N = 2$, fourth row: $n = 50$ and $K = 6$.

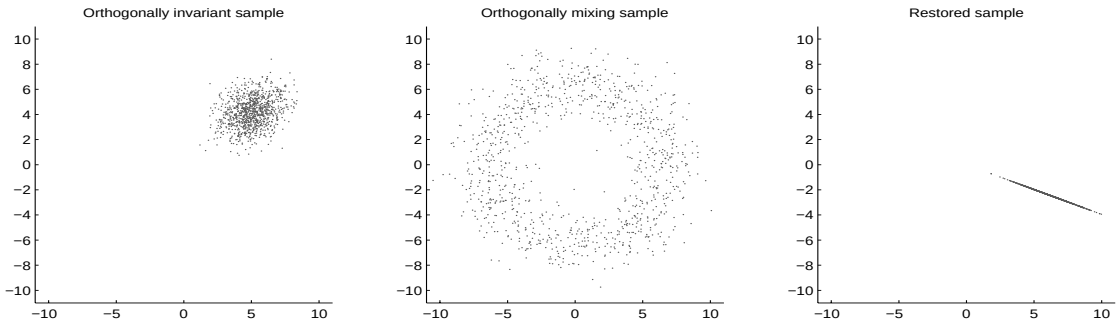

Figure 2.4: Sample from the orthogonally invariant distribution for $n = 1$ (left), orthogonally mixed sample (middle) and restored sample (right).

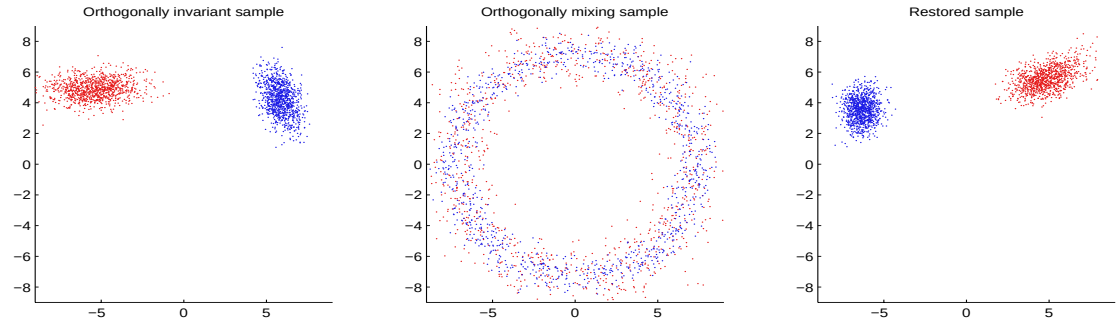

Figure 2.5: Sample from the orthogonally invariant distribution for $n = 2$ (left), orthogonally mixed sample (middle) and restored sample (right).

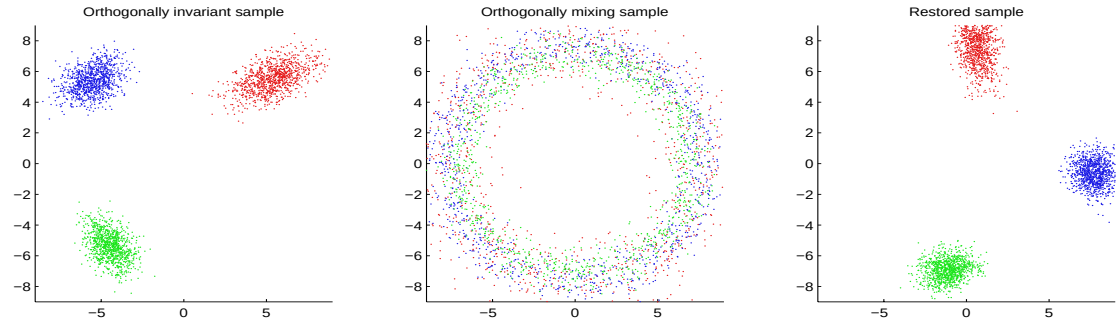

Figure 2.6: Sample from the orthogonally invariant distribution for $n = 3$ (left), orthogonally mixed sample (middle) and restored sample (right).

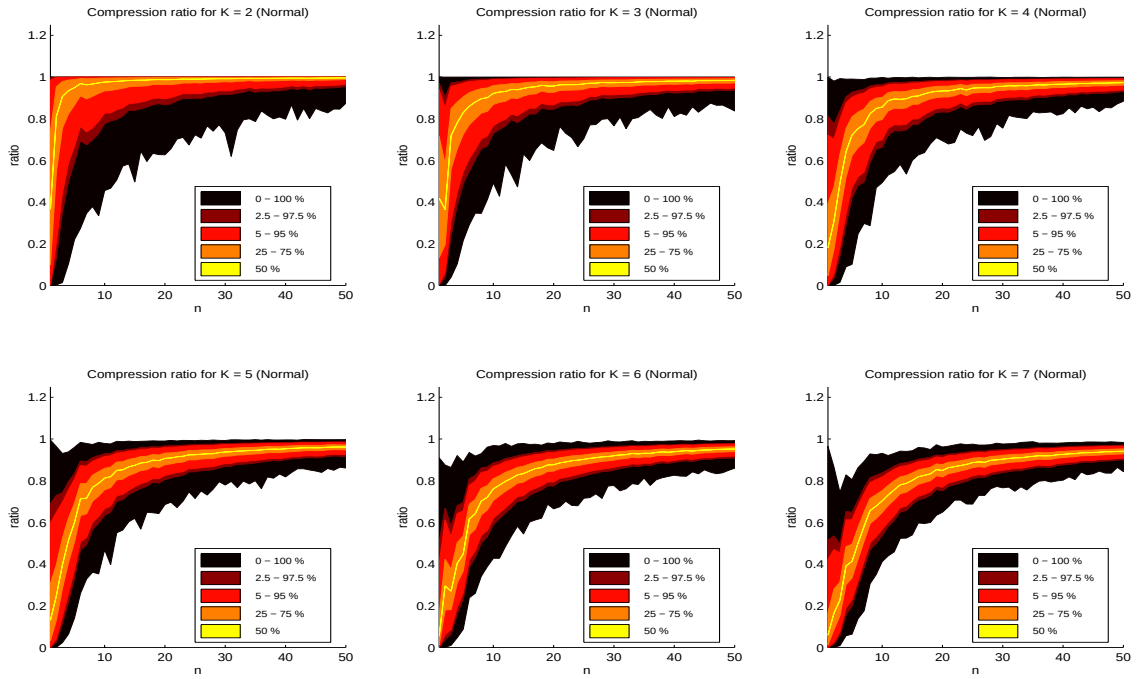

Figure 2.7: Quantiles of the error ratio for normally distributed data with known $\tilde{M}$.

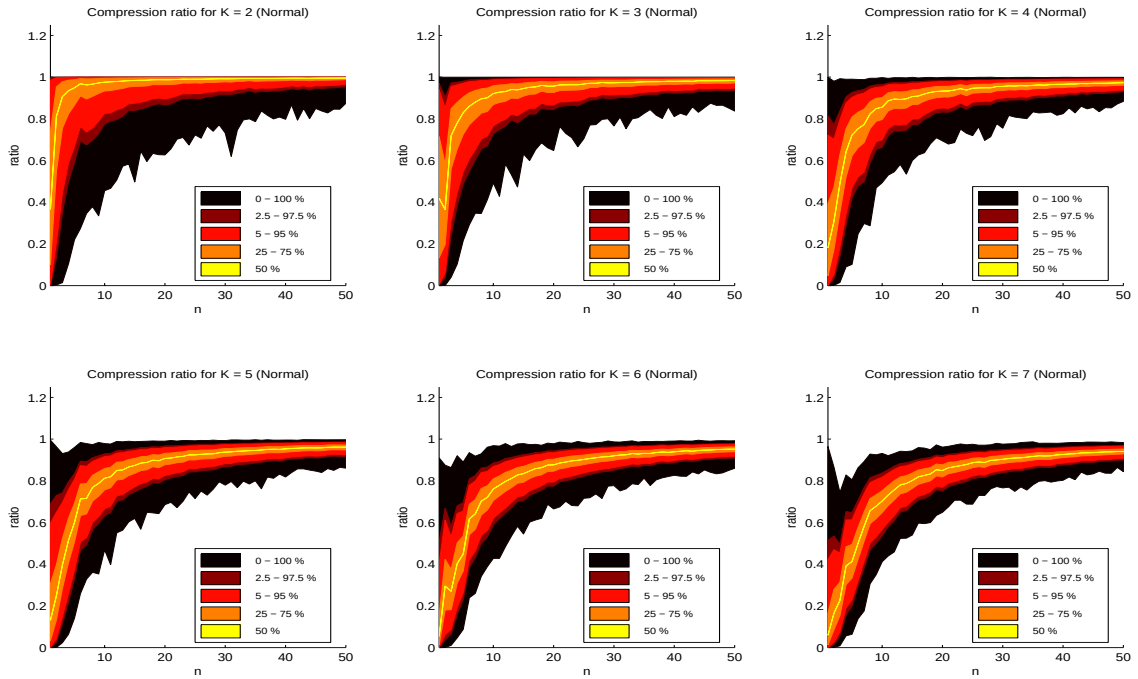

Figure 2.8: Quantiles of the error ratio for normally distributed data with unknown $\tilde{M}$.

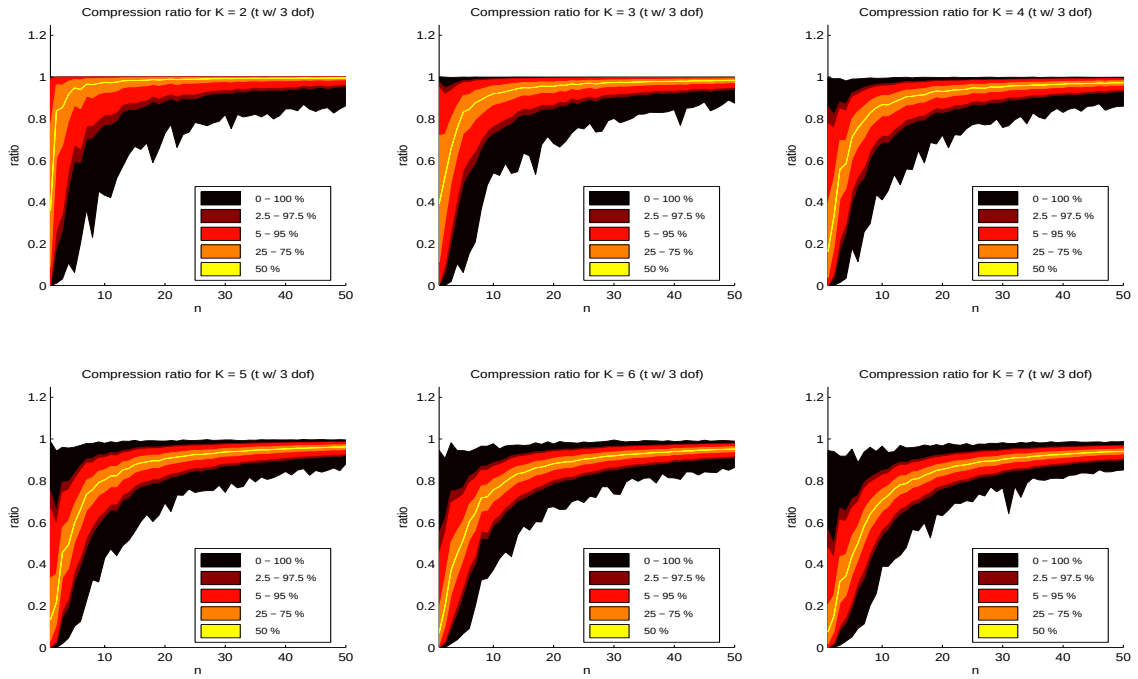

Figure 2.9: Quantiles of the error ratio for Student t distributed data with $\nu = 3$ and known $\tilde{M}$.

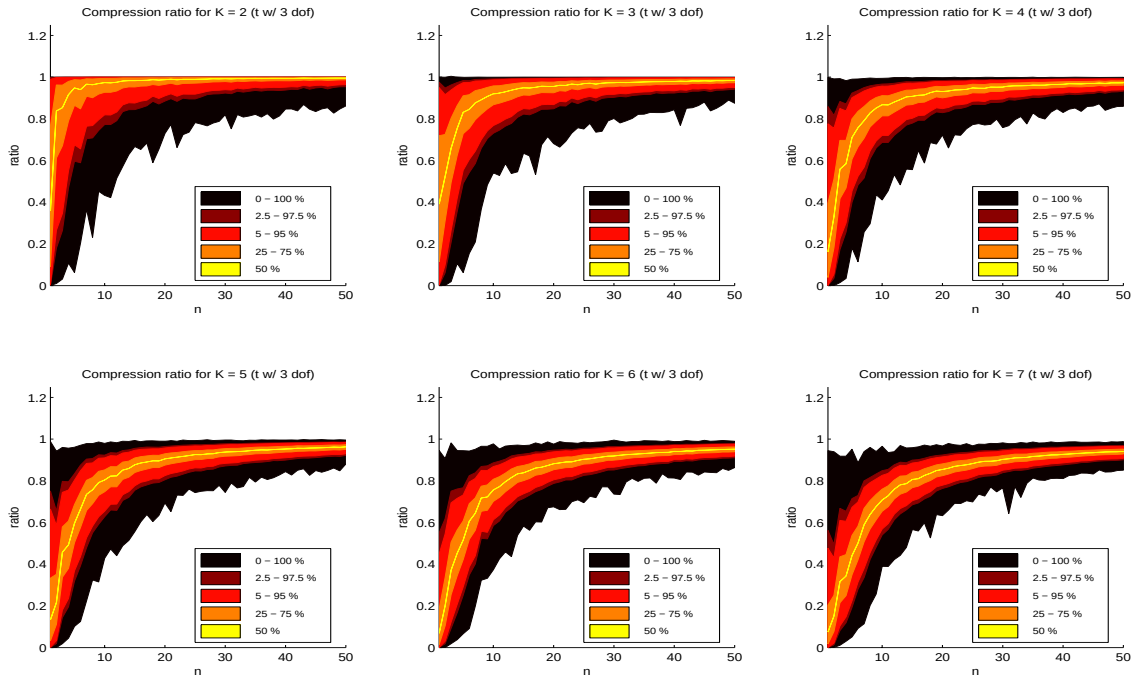

Figure 2.10: Quantiles of the error ratio for Student t distributed data with $\nu = 3$ and unknown $\tilde{M}$.

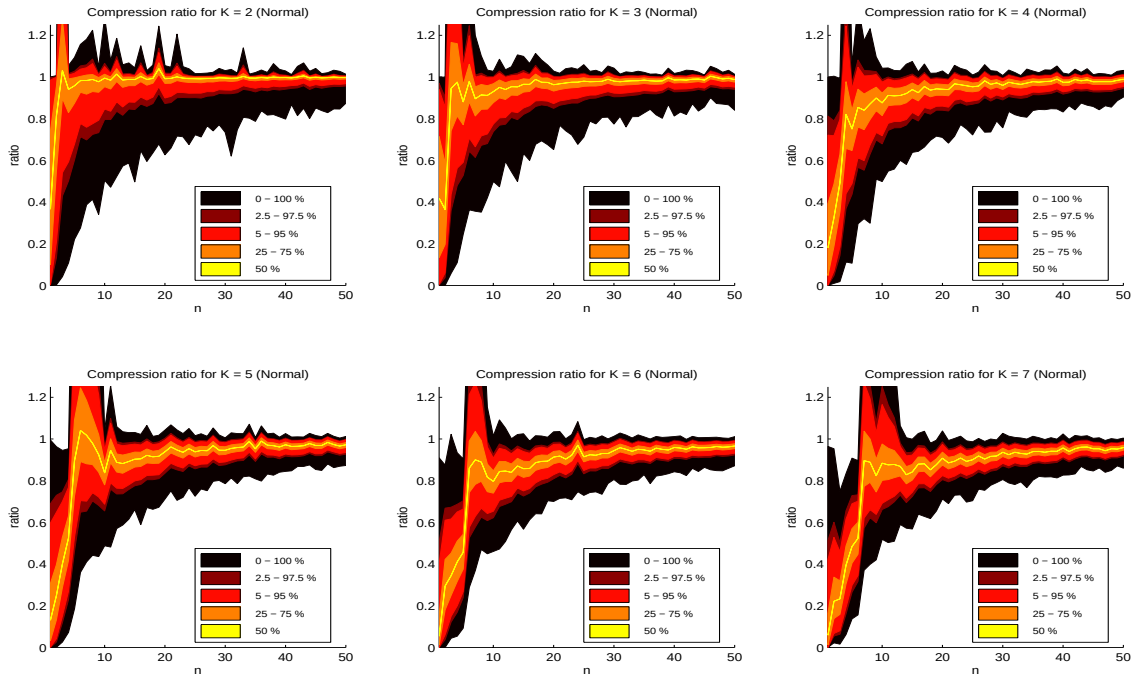

Figure 2.11: Quantiles of the error ratio for normally distributed data with unknown $\tilde{M}$, using the first weighting scheme.

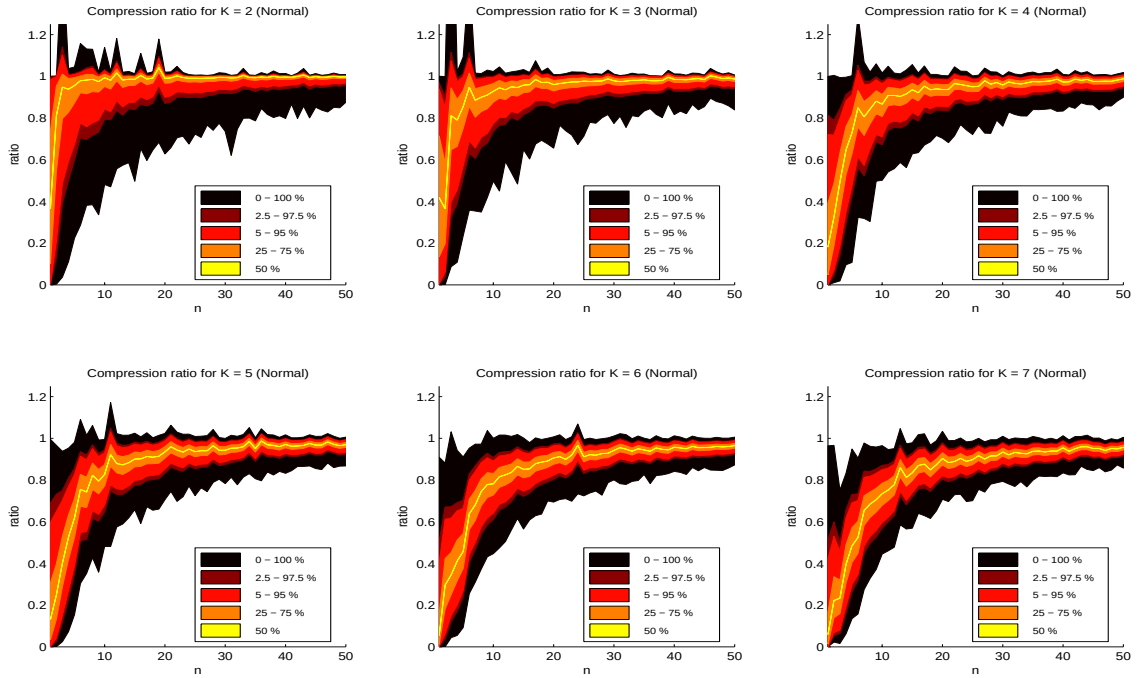

Figure 2.12: Quantiles of the error ratio for normally distributed data with unknown $\tilde{M}$, using the second weighting scheme.

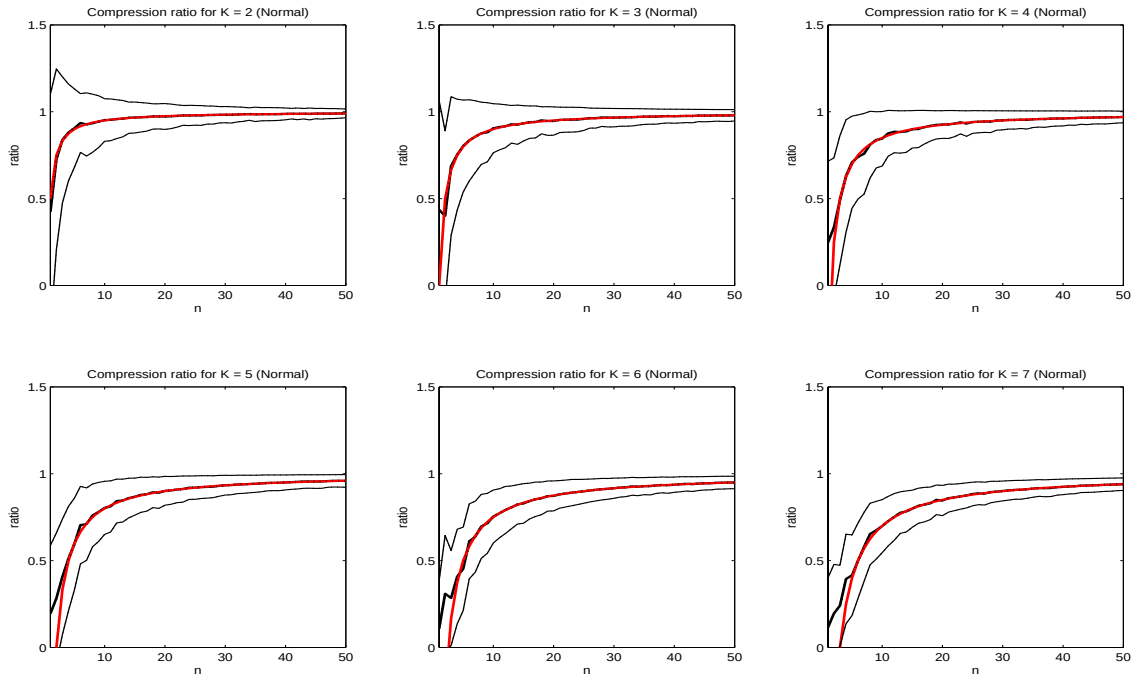

Figure 2.13: Mean of the error ratio ±2 standard deviations for normally distributed data.

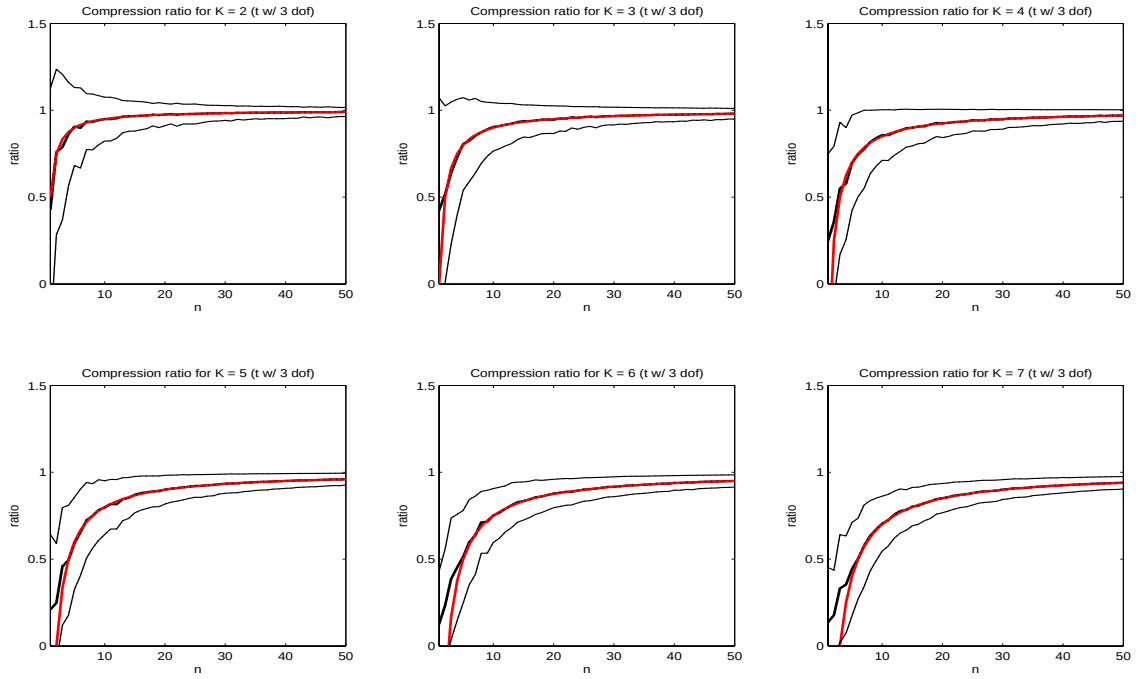

Figure 2.14: Mean of the error ratio ±2 standard deviations for Student t distributed data with $\nu = 3$.

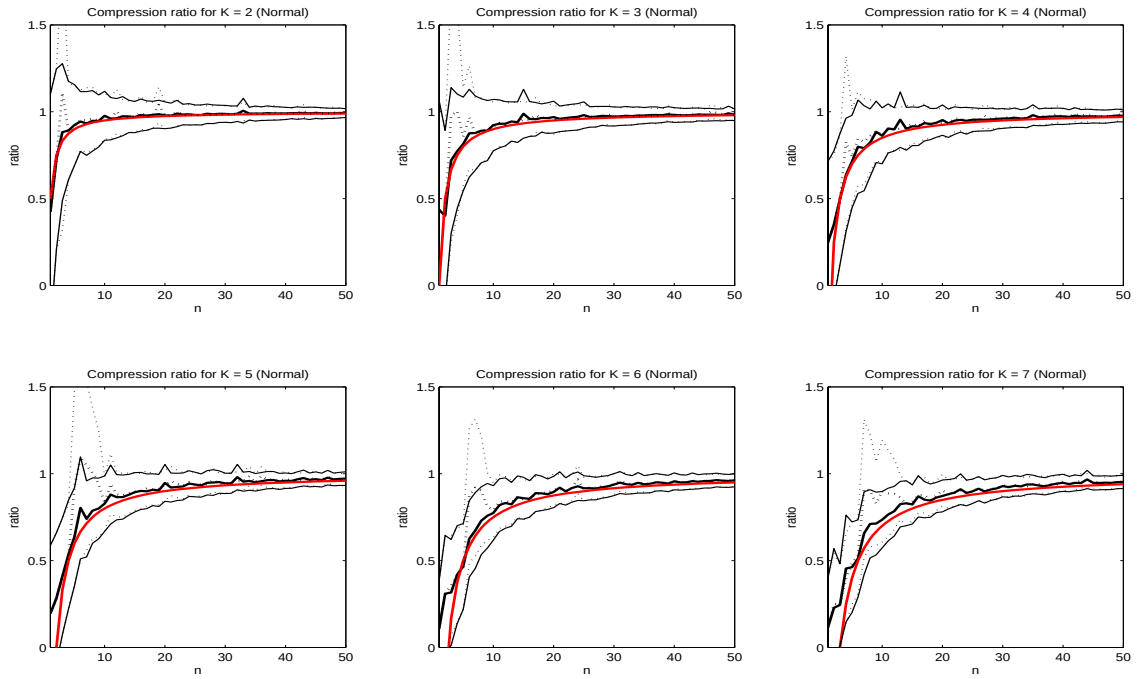

Figure 2.15: Mean of the error ratio ±2 standard deviations for normally distributed data, first weighting scheme.

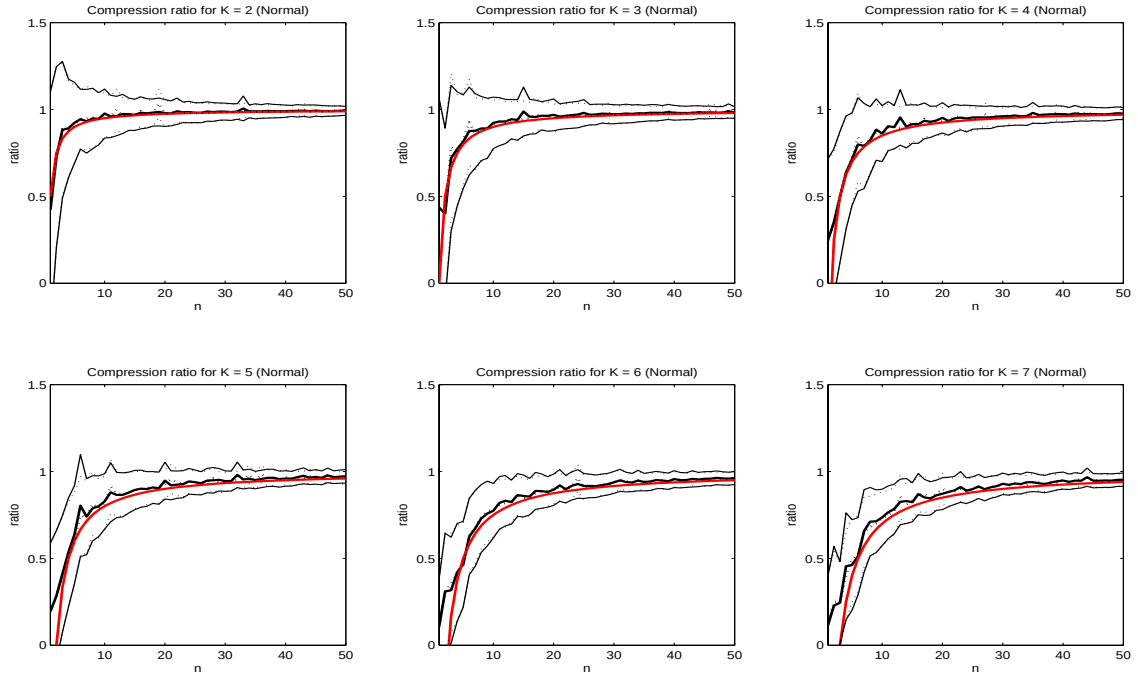

Figure 2.16: Mean of the error ratio ±2 standard deviations for normally distributed data, second weighting scheme.

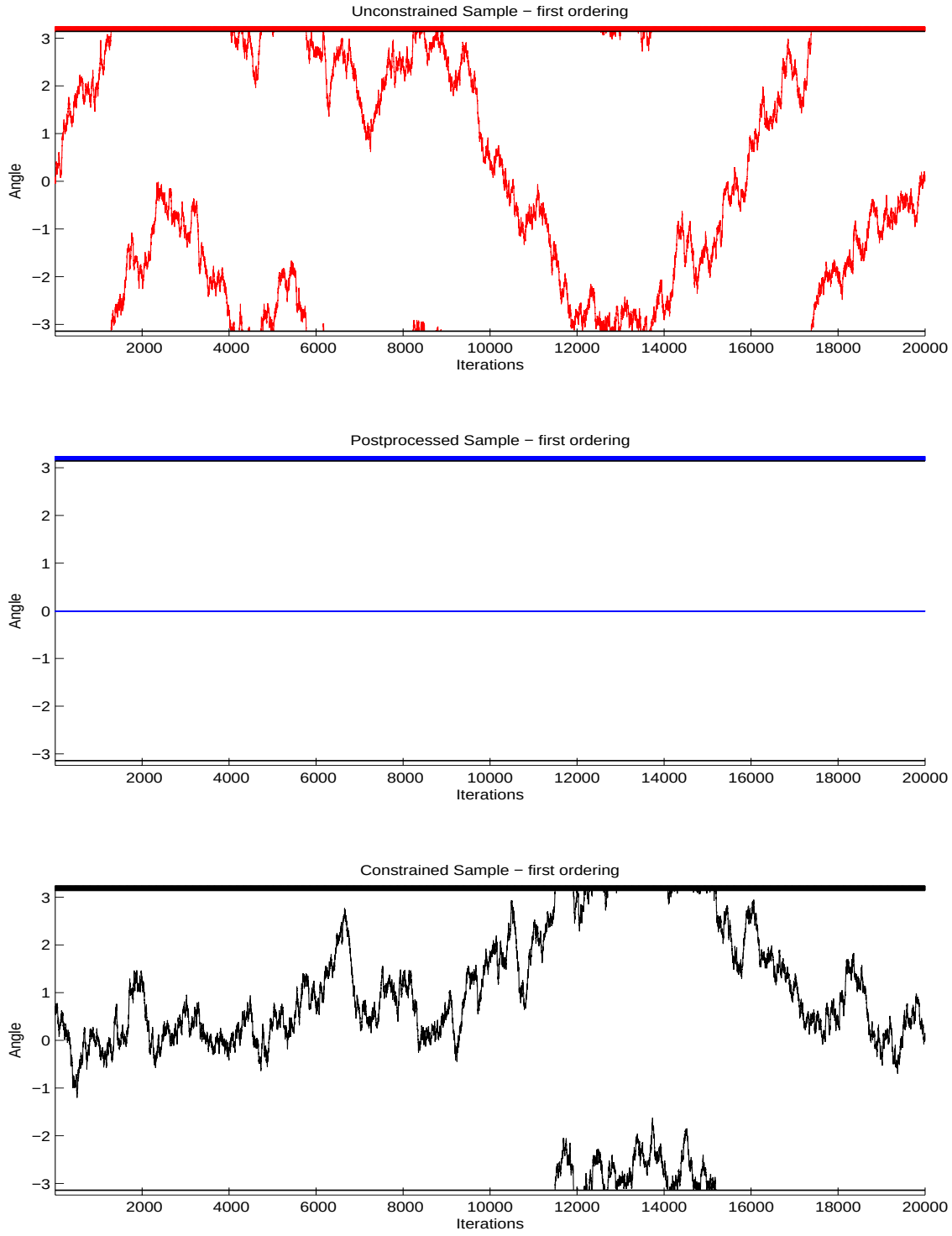

Figure 2.17: Estimated angles $\hat{\gamma}^{(z)}$ and reflection parameters $\hat{r}^{(z)}$ under the first ordering of the data

Notes: Angles and reflection parameters obtained from the orthogonal matrices from Algorithm 2.6.1, plotted for $20,000$ iterations after a burn-in sequence of $20,000$ iterations, using variables 1 and 2 serving as factor founders. First plot shows the results from the unconstrained sampler, second plot shows the postprocessed results from the unconstrained sampler, third plot shows the results from the constrained sampler.

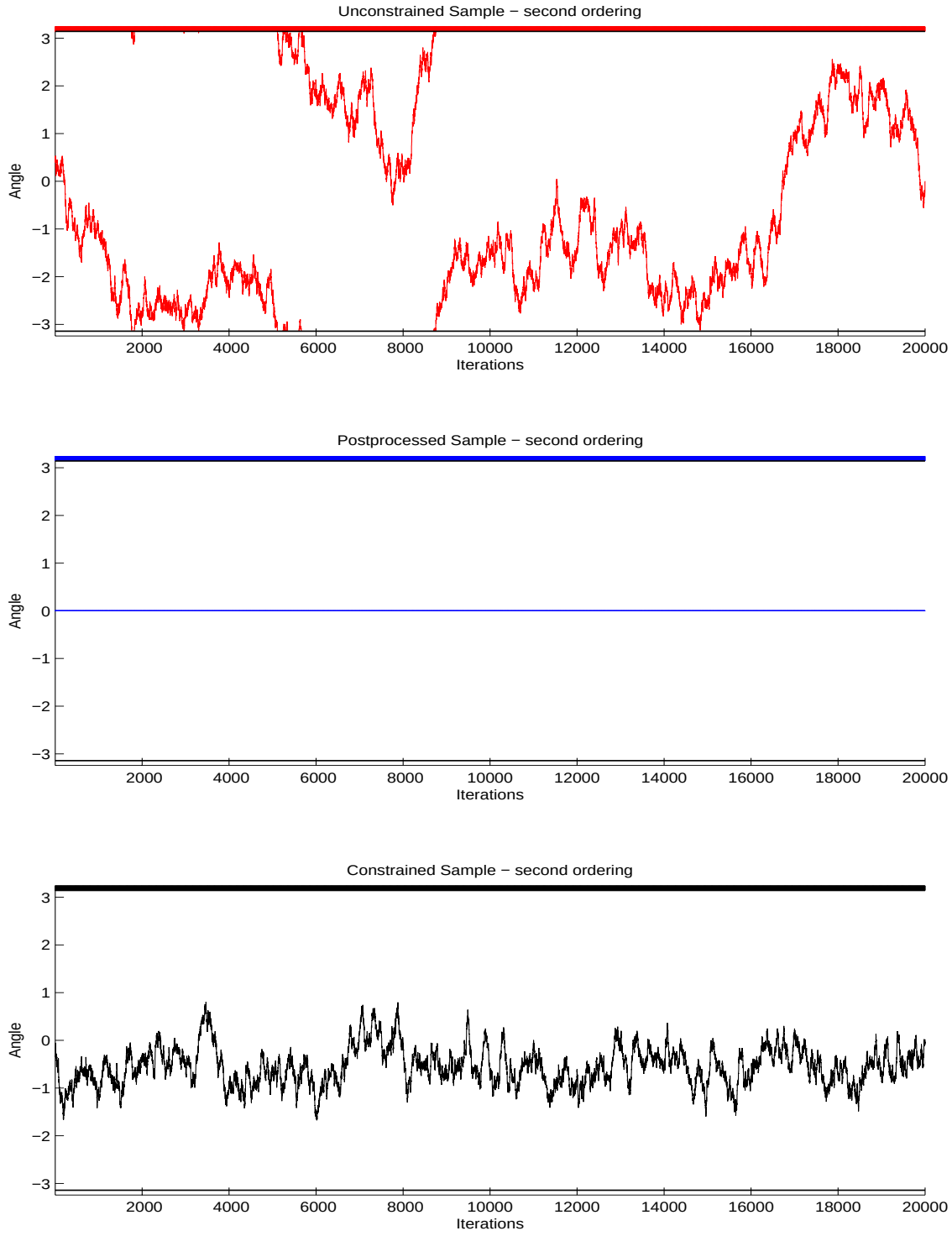

Figure 2.18: Estimated angles $\hat{\gamma}^{(z)}$ and reflection parameters $\hat{r}^{(z)}$ under the second ordering of the data.

Notes: Angles and reflection parameters obtained from the orthogonal matrices from Algorithm 2.6.1, plotted for $20,000$ iterations after a burn-in sequence of $20,000$ iterations, using variables 2 and 3 serving as factor founders. First plot shows the results from the unconstrained sampler, second plot shows the postprocessed results from the unconstrained sampler, third plot shows the results from the constrained sampler.

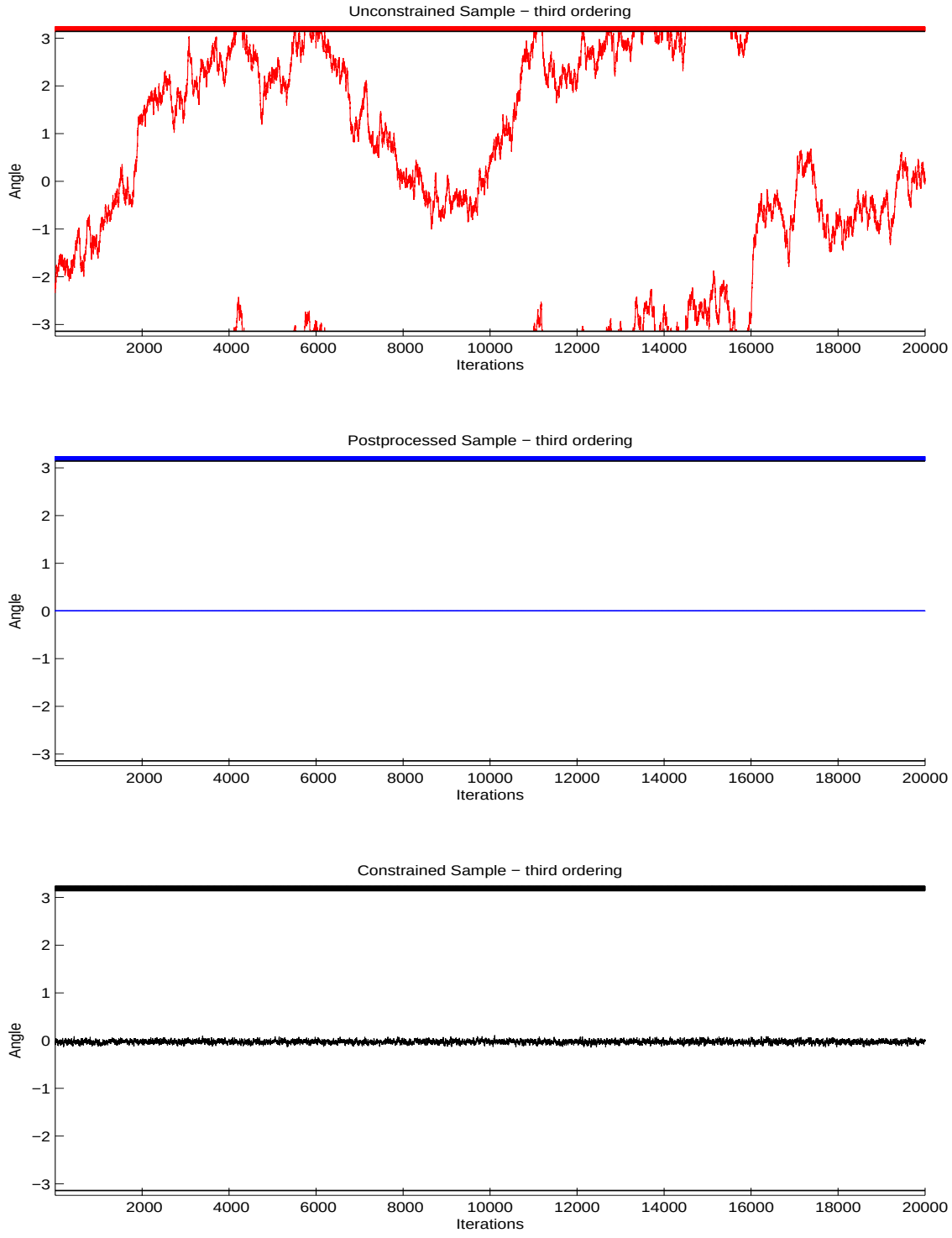

Figure 2.19: Estimated angles $\hat{\gamma}^{(z)}$ and reflection parameters $\hat{r}^{(z)}$ under the third ordering of the data.

Notes: Angles and reflection parameters obtained from the orthogonal matrices from Algorithm 2.6.1, plotted for $20,000$ iterations after a burn-in sequence of $20,000$ iterations, using variables 5 and 2 serving as factor founders. First plot shows the results from the unconstrained sampler, second plot shows the postprocessed results from the unconstrained sampler, third plot shows the results from the constrained sampler.

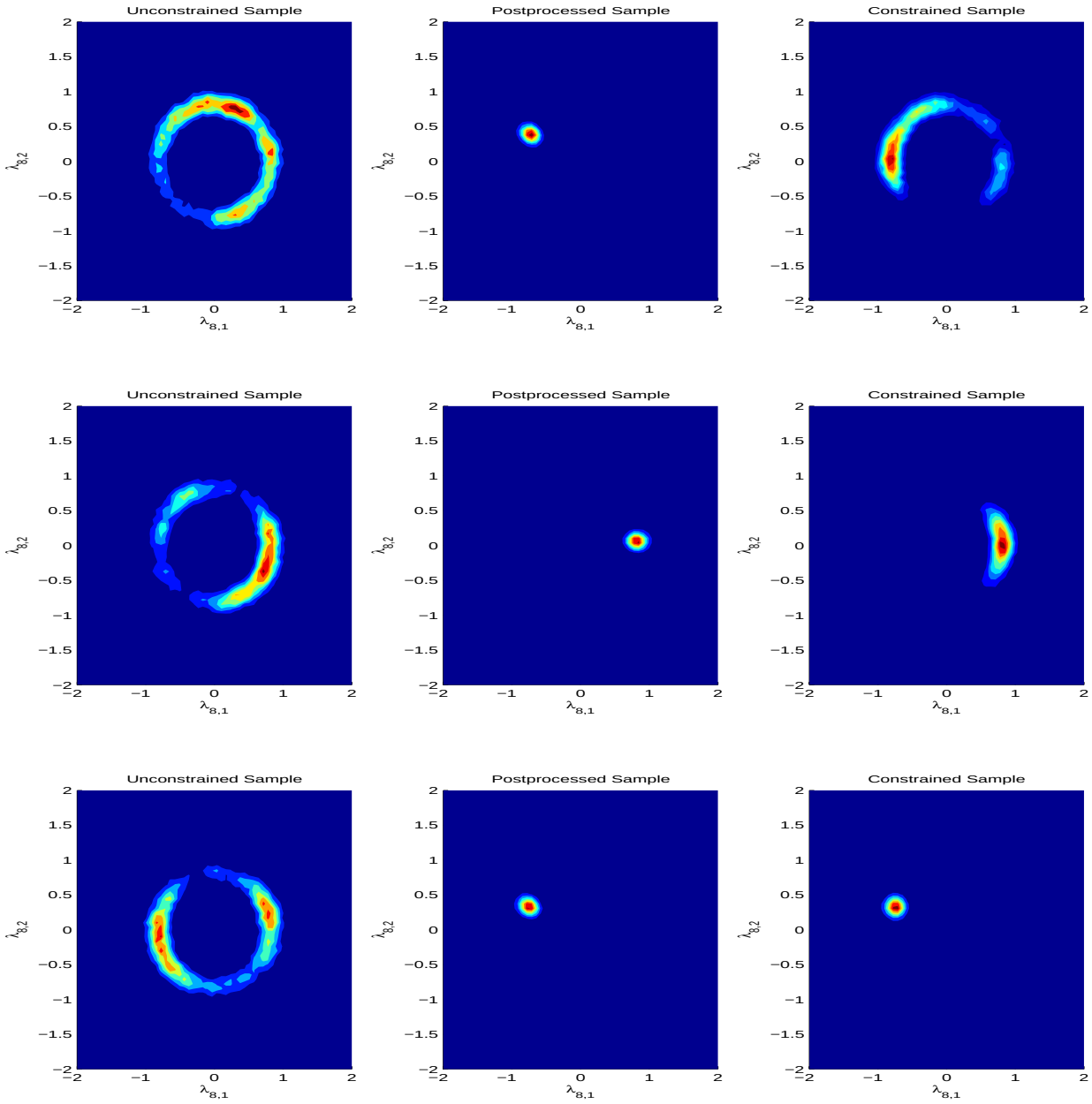

Figure 2.20: Contour plots displaying the orthogonal mixing resulting under the three orderings of the data.

Notes: First row shows the contour plots under the first ordering, second row shows the contour plots under the second ordering, and third row shows the contour plots under the third ordering. First column shows the results from the unconstrained sampler, second column shows the postprocessed results from the unconstrained sampler, and third column shows the results from the constrained sampler.

Appendix 2.A : Proof of Theorem 2.4.1

Proof. Assume $D \in \mathbb{R}^{K\times K}$. One way of performing a QR decomposition is to zero the $K(K-1)/2$ elements below the main diagonal. This can be obtained by means of a sequence of Givens rotations. Consider the $(2\times K)$ submatrices $D_{[i,j],[1:K]}$ for $i \in \{1,\dots,K\}$ and $j > i$. Those rotation matrices G_h that zero the matrix entry $d_{j,i}$ for each submatrix can be expressed in terms of an angle γ_h, where $h \in \{1,\dots,K(K-1)/2\}$. To obtain a unique solution for γ_h, let $d_{j,j} > 0$, i.e. if $d_{j,i} = 0$ and $d_{j,j} < 0$, add π to γ_h to perform a reflection about both axes. The result are two matrices $Q = \prod_{h=1}^{K(K-1)/2} G_h'$ with $Q \in \mathrm{SO}(K)$ and $R \in \mathbb{R}^{K\times K}$.

If $D \in \mathrm{SO}(K)$, $R \in \mathrm{SO}(K)$ with only positive elements on the diagonal, which can only hold if $R = I_K$. If $D \in \mathrm{O}(K)\backslash\mathrm{SO}(K)$, R must be an upper triangular matrix with positive elements on the diagonal except for $r_{K,K}$, which can only be -1, so R has $r_{i,j} = 0$ for all $i \neq j$, $r_{i,i} = 1$ for all $i < K$ and $r_{i,i} = -1$ for $i = K$. Hence, D can be expressed in terms of $\{\gamma_h\}_{h=1}^{K(K-1)/2}$ for $\gamma_h \in (-\pi,\pi]$ and $r_{K,K} \in \{-1,1\}$.

The according decomposition of an arbitrary orthogonal matrix D then looks as follows: The rotation part of D is expressed in terms of Givens rotation matrices G_1 to $G_{K(K-1)/2}$, where each matrix constitutes a rotation about a distinct pair of axes, or equivalently, of the corresponding columns of D. The matrices thus have the following form:

$$G_1 = \begin{pmatrix} \cos(\gamma_1) & -\sin(\gamma_1) & 0 & \dots & \dots & 0 \\ \sin(\gamma_1) & \cos(\gamma_1) & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \tag{2.69}$$

$$G_2 = \begin{pmatrix} \cos(\gamma_2) & 0 & -\sin(\gamma_2) & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & \dots & 0 \\ \sin(\gamma_2) & 0 & \cos(\gamma_2) & 0 & \dots & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \tag{2.70}$$

and so forth. Generally,

$$G_h = C_h \begin{pmatrix} \cos(\gamma_h) & -\sin(\gamma_h) \\ \sin(\gamma_h) & \cos(\gamma_h) \end{pmatrix} C_h', \tag{2.71}$$

where C_h is a $K \times 2$ matrix, with row vectors $c_{i_{h_1}} = [1\ \ 0]$ and $c_{i_{h_2}} = [0\ \ 1]$ and $c_i = [0\ \ 0]$ for $i \neq i_{h_1}, i \neq i_{h_2}$. i_{h_1} and i_{h_2} are the elements of the h^{th} two-element subset of $\{1,\dots,K\}$. Note

that there are exactly $K(K-1)/2$ such subsets, each representing a distinct pair of columns of D.

The reflection part of D is expressed as

$$B = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & \det(D) \end{pmatrix}, \tag{2.72}$$

i.e. B is the identity matrix if $\det(D) = 1$, and B is a reflection about the K^{th} axis if $\det(D) = -1$. □

Appendix 2.B : The Orthogonal Procrustes Algorithm

The orthogonal Procrustes procedure seeks to minimize the squared distance between the orthogonally transformed X_s and the target M, which is equivalent to minimizing the trace $\text{tr}(E'E)$ or the Frobenius norm $\|E'E\|_F$, where

$$E = X_s D - M \tag{2.73}$$

and minimizing

$$g = g_1 + g_2, \tag{2.74}$$

where

$$g_1 = \text{tr}(E'E) = \text{tr}(D'X_s'X_sD - 2D'X_s'M + M'M) \quad \text{and} \quad g_2 = \text{tr}(\Lambda(D'D - I_K)), \tag{2.75}$$

where Λ is a $K \times K$ matrix of Lagrange multipliers.

Taking the derivative with respect to D yields

$$\frac{\partial g}{\partial D} = (X_s'X_s + X_s'X_s)D - 2X_s'M + D(\Lambda + \Lambda') \stackrel{!}{=} 0. \tag{2.76}$$

Next, the terms are rearranged to obtain

$$\frac{\Lambda + \Lambda'}{2} = D'X_s'M - D'(X_s'X_s)D, \tag{2.77}$$

where the term on the left hand side and $D'(X_s'X_s)D$ are symmetric, so $D'X_s'M$ must be symmetric as well, i.e.

$$D'X_s'M = M'X_sD, \tag{2.78}$$

or, equivalently,

$$X_s'M = D'M'X_sD. \tag{2.79}$$

Taking the square of the latter yields

$$X_s'MM'X_s = D'M'X_sDD'X_s'MD = D'MX_sX_s'MD, \tag{2.80}$$

since $DD' = I$. A spectral decomposition on both $X_s'MM'X_s$ and $M'X_sX_s'M$ results in

$$WAW' = DVAV'D', \tag{2.81}$$

where the matrix of eigenvalues A is the same for both decompositions, while the eigenvectors V and W are different. Now

$$W = DV, \tag{2.82}$$

and, consequently,

$$D = WV'. \tag{2.83}$$

This is thus a necessary condition for D to be an orthogonal projection of X_s onto M. A minimum is thus also a minimum of

$$\begin{aligned} \text{tr}(E'E) &= \text{tr}(D'X'XD - 2D'X'M + M'M) \\ &= \text{tr}(X_s'X_s + M'M) - 2\text{tr}(D'X_s'M), \end{aligned} \tag{2.84}$$

which corresponds to a maximum of the second term, the first being fixed.

Plugging in the orthogonal projection solution for D from Equation (2.83) and the singular value decomposition

$$X_s'M = WA^{0.5}V' \tag{2.85}$$

yields

$$\begin{aligned} \text{tr}(D'X_s'M) &= \text{tr}(VW'X_s'M) \\ &= \text{tr}(VW'WA^{0.5}V') \\ &= \text{tr}(WW'A^{0.5}V'V) \\ &= \text{tr}(A^{0.5}), \end{aligned} \tag{2.86}$$

such that the singular value decomposition in Equation (2.85) yields the required matrices W and V that are needed in Equation (2.83), see also Golub and van Loan (2013), Algorithm 6.4.1.

Chapter 3

The Weighted Orthogonal Procrustes Approach for Static and Dynamic Factor Models

A version of this chapter dealing with the static factor model only has been published as CAU Kiel Economics Working Paper 2012-11 and Institute for the World Economy Kiel Working Paper 1799. An earlier version of this chapter has been published as Institute for the World Economy Kiel Working Paper 1902 and has been submitted for publication to the *Journal of Econometrics* on February 19, 2014. The revised version, which is identical to this chapter except for notational adjustments and Table 3.7 and references to this table, has been resubmitted on December 9, 2014.

3.1 Introduction

A latent factor model describes the influence of unobservable factors on observable data through factor loadings. Recent contributions discuss factor models in various contexts, see among others Conti et al. (2014), Kaufmann and Schumacher (2013), and Boivin and Ng (2006). In the factor model specification of Anderson and Rubin (1956), identifying assumptions are required for the model quantities of interest, such as factors and loadings. Anderson and Rubin (1956) deal with the question of model identification and show that after restricting the covariance of factor innovations to the identity matrix the model is still invariant under orthogonal transformations of loadings and factors. Anderson and Rubin (1956) call this the *rotation problem*.

Following the setup of Anderson and Rubin (1956), Geweke and Zhou (1996) discuss the Bayesian analysis of a factor model and deal with the rotation problem by constraining the loadings matrix to a positive lower triangular matrix, see also West (2003), Carneiro et al. (2003), Lopes and West (2004), and Carvalho et al. (2008). Bai and Wang (2012) show that the PLT approach solves the rotation problem also in the dynamic factor model setup. The PLT approach guarantees a unique global mode of the likelihood underlying the posterior distribution. It does not, however, preclude the existence of local modes. The constraints influence the shape of the likelihood and thus the shape of the posterior distribution, as discussed by Loken (2005) and Conti et al. (2014). This is problematic since local modes can

negatively affect the convergence behavior of Markov Chain Monte Carlo (MCMC) sampling schemes used for estimation purposes, see e.g. Celeux et al. (2000). As the constraints are imposed on particular elements of the loadings matrix, inference results may depend on the ordering of the variables. This is likewise observed by Carvalho et al. (2008). They call the variables whose loadings are constrained for identification purposes *factor founders* and develop an evolutionary search algorithm to choose the most appropriate subset of variables as factor founders. Similarly, Frühwirth-Schnatter and Lopes (2012) suggest a flexible approach that imposes a generalized lower triangular structure on the loadings matrix. Altogether, the use of ex-ante identification via constraints on the parameter space may influence inference results with respect to the model parameters and functions of these parameters.[1]

The use of parameter constraints for identification and their consequences on inference are also discussed in the econometric literature for finite mixture models. Similar to factor models, finite mixture models are typically not identified, as labels of the mixture components can be changed by permutation. Thus, given a permutation invariant prior distribution, the posterior distribution of finite mixture models has multiple symmetric modes. Identification can be achieved by fixing the ordering of the labels with respect to at least one of the parameters that are subject to label switching. However, if this identifying assumption is introduced by prior distributions, the choice of the constraint may have a substantial impact on the shape of the posterior distribution and estimates derived therefrom, see e.g. Stephens (2000). Moreover, the posterior distribution may have multiple local modes, which has severe consequences for the mixing behavior of the Gibbs sampler. To cure this problem in the context of finite mixtures Celeux (1998) and Stephens (2000) suggest to achieve identification via postprocessing the output of the unconstrained sampler using relabeling algorithms.

In correspondence to the literature on finite mixture models, we propose an ex-post approach to fix the rotation problem that is suitable for the Bayesian analysis of both static and dynamic factor models. The proposed ex-post approach towards the rotation problem can be framed as a decision theoretic approach, compare Celeux (1998), Celeux et al. (2000), and Stephens (2000). The suggested approach does not constrain the parameter space, but fixes the rotation problem by re-transforming the output of the unconstrained Gibbs sampler using a sequence of orthogonal matrices. This sequence of orthogonal matrices is determined using a loss function adequately defined for the static and dynamic factor model. The minimization of the corresponding expected loss in static factor models is based on the orthogonal Procrustes transformation proposed by Kristof (1964) and Schönemann (1966). Additionally, a weighting scheme as discussed by Lissitz et al. (1976) can be used, hence we refer to the suggested approach as the weighted orthogonal Procrustes (WOP) approach. For the dynamic factor model we use a parametrization of orthogonal matrices allowing for numerical minimization of the defined expected loss.[2] The suggested ex-post WOP approach

[1] Accordingly, Lopes and West (2004) find that model selection criteria used to choose the number of factors are influenced by the way the variables are ordered and thus by the position of the restrictions on the parameter space.

[2] In an approach for sparse factor models, Kaufmann and Schumacher (2013) perform temporary orthogonal transformations of the model parameters to satisfy an alternative identification scheme suggested by

towards the rotation problem provides order invariant inference, since any permutation of the variables invokes the same expected loss. This is in contrast to the ex-ante PLT approach, where differences between inference obtained under different orderings cannot be attributed to a single orthogonal transformation only. Further, the ex-post WOP approach allows for transforming the obtained estimators via a single orthogonal matrix as implied by criteria like Varimax or Quartimax to enhance interpretability. In turn, interpretational assumptions do not interfere with estimation. In this sense, the ex-post WOP approach is purely exploratory.

To illustrate the properties of our ex-post WOP approach, we provide a simulation study with static and dynamic factor models. We compare our inference results from the ex-post WOP approach with those from the ex-ante PLT approach by Geweke and Zhou (1996). We check both corresponding samplers for their convergence properties, as well as statistical and numerical accuracy. Convergence is generally obtained faster for the WOP approach. In some of the considered scenarios, the PLT approach provides a multimodal and highly skewed posterior distribution for the loading parameters. Across all considered model setups and prior scenarios, this is not observed for the WOP approach, which also shows much higher numerical accuracy than the PLT approach.

In an empirical application, we analyze the panel of 120 macroeconomic time series from Bernanke et al. (2005) using both the PLT approach and the WOP approach. As the first exercise, we choose series as factor founders that are particularly fit for this purpose and estimate the model repeatedly. Afterwards, we perform repeated estimations of the model under randomly chosen orderings of the series. The WOP approach is found to be numerically more stable than the PLT approach in both exercises.

The paper proceeds as follows. Section 3.2 provides the dynamic factor model and briefly discusses the identification of the model. Section 3.3 introduces the novel ex-post approach towards the rotation problem for static and dynamic factor models. Section 3.4 illustrates the differences between the WOP approach and the PLT approach by means of a simple example. Section 3.5 presents a simulation study that compares both approaches. Section 3.6 provides an empirical illustration using the data set of Bernanke et al. (2005). Section 3.7 concludes.

3.2 Model Setup, Identification, and the Rotation Problem

In a dynamic factor model the comovements in a data panel with N variables and time dimension T are represented by K factors that relate to the data via loadings. The dynamic factor model takes the form

$$y_t = \Lambda_0 f_t + \Lambda_1 f_{t-1} + ... + \Lambda_S f_{t-S} + e_t, \quad t = 1, \ldots, T, \tag{3.1}$$

Anderson and Rubin (1956), such that the outer product of the loadings matrix with itself is diagonal with decreasingly ordered elements. Under this identification, the latent factors are sampled and afterwards transformed back into the original parametrization. This approach works well for sparse factor models, but seems to be inappropriate for the purely exploratory factor analysis as discussed here.

where y_t is an $N \times 1$ demeaned and stationary vector of observed data, f_t is a $K \times 1$ vector of K latent factors, Λ_s, $s = 0, \ldots, S$ representing $N \times K$ matrices of loadings, and e_t denotes a $N \times 1$ vector of errors with e_t being independently and identically normally distributed with mean zero and covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$.[3] Further, the K factors follow a vector autoregressive process of order P given as

$$f_t = \Phi_1 f_{t-1} + \Phi_2 f_{t-2} + ... + \Phi_P f_{t-P} + \epsilon_t, \tag{3.2}$$

where Φ_p, $p = 1, \ldots, P$ are $K \times K$ persistence matrices, and ϵ_t denotes the error being independently and identically normally distributed with mean zero and covariance equal to the K-dimensional identity matrix I_K. Setting the covariance of ϵ_t to the identity matrix solves the identification problem up to the rotation problem as discussed by Anderson and Rubin (1956) for static factor models. Bai and Wang (2012) show that this also holds for the dynamic factor model described here. We consider the likelihood with

$$\vartheta = (\text{vec}(\Lambda_0), \ldots, \text{vec}(\Lambda_S), \text{vec}(\Phi_1), \ldots, \text{vec}(\Phi_P), \text{diag}(\Sigma)) \tag{3.3}$$

summarizing all model parameters, $Y = (y_1, \ldots, y_T)$ and $f_0 = \ldots = f_{-\max\{S-1, P-1\}} = 0$ given as

$$\begin{aligned}
\mathcal{L}(Y|\vartheta) &= \int_{f_T} \cdots \int_{f_1} \prod_{t=1}^{T} p(y_t|\vartheta, f_t, \ldots, f_{t-S}) p(f_t|\vartheta, f_{t-1}, \ldots, f_{t-P}) df_1 \ldots df_T \qquad (3.4)\\
&= \int_{f_T} \cdots \int_{f_1} (2\pi)^{-\frac{TN}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \left((y_t - \sum_{s=0}^{S} \Lambda_s f_{t-s})' \Sigma^{-1} (y_t - \sum_{s=0}^{S} \Lambda_s f_{t-s}) \right) \right\} \\
&\quad (2\pi)^{-\frac{TK}{2}} |I_K|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (f_t - \sum_{p=1}^{P} \Phi_p f_{t-p})' (f_t - \sum_{p=1}^{P} \Phi_p f_{t-p}) \right\} df_1 \ldots df_T.
\end{aligned}$$

The likelihood is invariant under the following parameter transformation.[4] Define for any orthogonal $K \times K$ matrix D the transformation

$$H(D)\vartheta = (\text{vec}(\Lambda_0 D), \ldots, \text{vec}(\Lambda_S D), \text{vec}(D'\Phi_1 D), \ldots, \text{vec}(D'\Phi_P D), \text{diag}(\Sigma)), \tag{3.5}$$

with

$$H(D) = \begin{pmatrix} (D' \otimes I_{N(S+1)}) & 0 & 0 \\ 0 & I_P \otimes (D' \otimes D') & 0 \\ 0 & 0 & I_N \end{pmatrix}, \tag{3.6}$$

[3] Note that the model could be further extended by an autoregressive process of order Q for errors e_t as discussed by Kaufmann and Schumacher (2013).

[4] The likelihood is also invariant under any permutation of the variables in y_t, as well as the rows of each Λ_s, $s = 0, \ldots, S$ and the corresponding diagonal elements of Σ. This order invariance is not present in the PLT approach placing constraints on the loading parameters.

where $|\det(H^{-1}(D))| = 1$. For completion, considering $\tilde{f}_t = D'f_t$, $t = 1, \ldots, T$, and $d\tilde{f}_t = |\det(D)|df_t = df_t$ and taking into account that the transformation has no impact on the range of parameters yields $\mathcal{L}(Y|\vartheta) = \mathcal{L}(Y|H(D)\vartheta)$, i.e. the likelihood remains the same under the transformation in Equation (3.5). We refer to this invariance of the likelihood as the rotation problem.[5]

The invariance of the likelihood transfers to the posterior distribution and thus the posterior estimators, when the chosen prior distribution is as well invariant under the transformation described in Equation (3.5). As the rotation problem does not involve Σ, we choose the commonly used conditional conjugate prior as independent inverse gamma distributions with probability density

$$\pi(\Sigma) = \prod_{i=1}^{N} \frac{\underline{\beta}_i^{\underline{\alpha}_i}}{\Gamma(\underline{\alpha}_i)} \sigma_i^{-2(\underline{\alpha}_i+1)} \exp\left\{-\frac{\underline{\beta}_i}{\sigma_i^2}\right\}. \tag{3.8}$$

To ensure the invariance to the orthogonal transformation as stated in Equation (3.5), the priors for $\bar{\Lambda} = (\Lambda_0', \ldots, \Lambda_S')'$ and Φ_p, $p = 1, \ldots, P$ are chosen as

$$\pi(\Phi_1, \ldots, \Phi_P) \propto c, \quad c > 0, \tag{3.9}$$

and

$$\pi(\Lambda_0, \ldots, \Lambda_S) = \prod_{s=0}^{S} (2\pi)^{-\frac{KN}{2}} |\Omega_{\Lambda_s}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\text{vec}(\Lambda_s) - \mu_{\Lambda_s})'\Omega_{\Lambda_s}^{-1}(\text{vec}(\Lambda_s) - \mu_{\Lambda_s})\right\} \tag{3.10}$$

respectively. The normal prior for $\{\Lambda_s\}_{s=0}^{S}$ is in line with the specification of Bai and Wang (2012), but does not impose constraints. The constant prior for Φ_p, $p = 1, \ldots, P$ likewise follows the specification of Bai and Wang (2012), or, more generally, the specifications for Bayesian vector autoregressive modeling by Ni and Sun (2005). Additional stationarity constraints can be imposed by demanding that the eigenvalues of the companion matrix of $\{\Phi_p\}_{p=1}^{P}$ are all less than 1 in absolute value, see e.g. Hamilton (1994), Chapter 10. Note that the eigenvalues of the companion matrix are unaffected by the transformation in Equation (3.5).[6] We require that all mean vectors are set to zero, i.e. $\mu_{\Lambda_s} = 0$, $s = 0, \ldots, S$ and $\Omega_{\Lambda_s} = \underline{\Upsilon}_s \otimes I_K$, $s = 0, \ldots, S$ where each $\underline{\Upsilon}_s$ is a positive diagonal $N \times N$ matrix. The so far stated posterior distribution

$$p(\vartheta|Y) \propto \mathcal{L}(Y|\vartheta)\pi(\Sigma)\pi(\Phi_1, \ldots, \Phi_P)\pi(\Lambda_0, \ldots, \Lambda_S) \tag{3.11}$$

[5] The static case arising for $S = P = 0$ corresponds to the closed form likelihood given as

$$(2\pi)^{-\frac{TN}{2}} |\Lambda_0\Lambda_0' + \Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} y_t'(\Lambda_0\Lambda_0' + \Sigma)^{-1}y_t\right\}. \tag{3.7}$$

With regard to invariance of the likelihood, the same caveats as in the dynamic case apply.

[6] The constant prior for Φ_p, $p = 1, \ldots, P$ can also be replaced by normal priors with zero mean and a covariance matrix that equals the identity matrix times a constant, since this distribution is also not affected by the transformation in Equation (3.5).

is then invariant under the transformation in Equation (3.5).

The model setup is directly accessible in state-space form. This allows for sampling using the methodology presented in Carter and Kohn (1994). Appendix 3.A gives a detailed description of the corresponding Gibbs sampler, which we call the *unconstrained Gibbs sampler* in the following, because it does not impose any constraints on the loadings matrix in order to solve the rotation problem. The output of the unconstrained sampler allows for conducting inference on rotation invariant quantities such as the variance of the idiosyncratic errors Σ and the systematic part $\sum_{s=0}^{S} \Lambda_s f_{t-s}$ and thus also $y_t - \sum_{s=0}^{S} \Lambda_s f_{t-s}$. This is particularly important since methods exist to determine the number of factors based on these rotation invariant quantities, see Chan et al. (2013). Typically used criteria to determine the number of factors in frequentist setups are e.g. described in Bai and Ng (2002, 2007), Breitung and Pigorsch (2013) or Onatski (2010). These approaches are based on rotation invariant quantities and could be adopted to the output of the unconstrained sampler. The ex-post WOP approach is meant to be applied after the number of factors has been determined.

3.3 An Ex-Post Approach Towards the Rotation Problem

The rotation problem is solved when the uniqueness of the parameter estimator derived from the posterior distribution is ensured. The uniqueness is ensured when the invariance of the posterior distribution under the transformation in Equation (3.5) is inhibited. This is possible via ex-ante restrictions on the parameter space hindering the mapping of any points within the admissible parameter space by orthogonal matrices. While ex-ante restrictions are routinely applied in many econometric frameworks, ex-post identification is prominent for finite mixture models, see Celeux et al. (2000), Stephens (2000), Frühwirth-Schnatter (2001); Frühwirth-Schnatter (2006) and Grün and Leisch (2009).[7] To address the rotation problem, we propose an ex-post approach for Bayesian analysis of static and dynamic factor models, which can also be motivated as a decision-theoretic approach, see e.g. Stephens (2000). A decision-theoretic approach uses a loss function to assess the difference between the parameter taking value ϑ and the corresponding estimator $\hat{\vartheta}$, where one possibility to operationalize this difference is the quadratic distance. Following Jasra et al. (2005), a loss function $L(\hat{\vartheta}, \vartheta)$ is defined as a mapping of the estimators $\hat{\vartheta}$ from the set of possible estimators Ξ and each of the parameter

[7] In the context of finite mixture models ex-post identification is used as the posterior is invariant under permutation of mixing components, i.e. when according to Redner and Walker (1984) label switching occurs in the output of an unconstrained sampler. Richardson and Green (1997) advise to use different identifiability constraints when postprocessing the MCMC output. Stephens (2000) and Frühwirth-Schnatter (2001) propose the use of relabeling algorithms that screen the output of the unconstrained sampler and sort the labels to minimize some divergence measures, e.g. Kullback-Leibler distances. The main idea behind the relabeling approach in finite mixtures is that the output of the unconstrained sampler in fact stems from a mixture distribution. The mixing is discrete and occurs via permutations of the labels. The relabeling algorithm fixes the invariance of the likelihood with respect to a specific permutation based on a decision criterion and reverses thus the mixing.

values ϑ within the parameter space on the real line, i.e. $L : \Xi \times \vartheta \to [0, \infty)$. The optimal estimator in terms of minimal expected loss is then defined as

$$\tilde{\hat{\vartheta}} = \arg\min_{\hat{\vartheta}} \int_{\vartheta} L(\hat{\vartheta}, \vartheta) p(\vartheta|Y) d\vartheta. \tag{3.12}$$

For computational reasons, a Monte Carlo (MC) approximation is used for the integral involved in Equation (3.12), thus we obtain

$$\tilde{\vartheta}^* = \arg\min_{\vartheta^*} \frac{1}{Z} \sum_{z=1}^{Z} L(\vartheta^*, \vartheta^{(z)}), \tag{3.13}$$

where $\vartheta^{(z)}$, $z = 1, \ldots, Z$ denotes a sample from the unconstrained posterior distribution and ϑ^* the MC analog to $\hat{\vartheta}$.

The suggested ex-post WOP approach is based on the observation that the unconstrained sampler provides a realized sample $\{\vartheta^{(z)}\}_{z=1}^{Z}$ from the posterior distribution which can equivalently be interpreted as a sample taking the form $\{H(D^{(z)})\vartheta^{(z)}\}_{z=1}^{Z}$, i.e. a sample given as a transformation of the realized sample by an arbitrary sequence of orthogonal matrices $\{D^{(z)}\}_{z=1}^{Z}$. All samples taking the form $\{H(D^{(z)})\vartheta^{(z)}\}_{z=1}^{Z}$ are assigned the same posterior probability. Due to this indeterminacy, we refer to the unconstrained sample as *orthogonally mixed*. Each choice for the sequence $\{D^{(z)}\}_{z=1}^{Z}$ results in a different estimate of ϑ. To distinguish between the different possible forms $\{H(D^{(z)})\vartheta^{(z)}\}_{z=1}^{Z}$, and correspondingly ensure uniqueness of the estimate, we advocate to extend the loss function approach in order to discriminate between the losses invoked under different orthogonal transformations of the realized sample. The resulting loss function then takes the form

$$L(\vartheta^*, \vartheta^{(z)}) = \min_{D^{(z)}} \{L_D(\vartheta^*, H(D^{(z)})\vartheta^{(z)})\}, \quad \text{s.t. } D^{(z)\prime} D^{(z)} = I_K, \tag{3.14}$$

with $L_D(\vartheta^*, H(D^{(z)})\vartheta^{(z)})$ denoting for given ϑ^* the loss invoked for any transformation of $\vartheta^{(z)}$ as described in Equation (3.5). If for any $\vartheta^{(z)}$ the minimal loss can be uniquely determined, *orthogonal mixing* is immaterial for parameter estimation and the rotation problem is fixed. The choice of the loss function allowing for discriminating different sequences of orthogonal matrices is restricted with regard to solvability and uniqueness of the solution to the particular minimization problem.[8] We suggest a quadratic loss function for the considered static and dynamic factor model denoted as

$$L_D(\vartheta^*, H(D^{(z)})\vartheta^{(z)}) = \operatorname{tr}\left[(H(D^{(z)})\vartheta^{(z)} - \vartheta^*)'(H(D^{(z)})\vartheta^{(z)} - \vartheta^*)\right]. \tag{3.15}$$

[8] In the finite mixture context, see e.g. Sperrin et al. (2010), it is typical to base the loss function on the Kullback-Leibler distance, also referred to as relative entropy, between the posterior distribution and the distribution of interest. Although the Kullback-Leibler distance has the desired properties, see e.g. Clarke et al. (1990), we opt for a quadratic loss function as first-order equivalent under general regularity conditions, see Cheng et al. (1999), for reasons of solvability.

Using the MC version of the expected posterior loss results in the following minimization problem

$$\{\{\tilde{D}^{(z)}\}_{z=1}^{Z}, \tilde{\vartheta}^*\} = \underset{\{D^{(z)}\}_{z=1}^{Z}, \vartheta^*}{\operatorname{argmin}} \sum_{z=1}^{Z} L_D(\vartheta^*, H(D^{(z)})\vartheta^{(z)}),$$
$$\text{s.t. } D^{(z)\prime}D^{(z)} = I_K, \; z = 1, \ldots, Z. \tag{3.16}$$

As this defined expected loss function is globally convex in ϑ^*, all minima can be characterized as

$$\frac{1}{Z}\sum_{z=1}^{Z} \operatorname{tr}\left[\left(H(D^{(z)})\vartheta^{(z)} - \overline{H(D)\vartheta}\right)'\left(H(D^{(z)})\vartheta^{(z)} - \overline{H(D)\vartheta}\right)\right] \tag{3.17}$$

with

$$\overline{H(D)\vartheta} = \frac{1}{Z}\sum_{z=1}^{Z} H(D^{(z)})\vartheta^{(z)}. \tag{3.18}$$

Therefore, for the sample as implied by $\{H(\tilde{D}^{(z)})\vartheta^{(z)}\}_{z=1}^{Z}$, the rotation problem is solved.

The following paragraphs outline how a solution for the static as well as the dynamic factor model setup can obtained via a sequential algorithm.

Static Factor Model A solution to the optimization problem stated in Equation (3.16) applied to the factor model with $P = 0$ is obtained iteratively via a two-step optimization. The algorithm needs an initialization with regard to $\vartheta^* = \{\operatorname{vec}(\bar{\Lambda}^*), \operatorname{diag}(\Sigma^*)\}$, where we choose the last draw of the unconstrained sampler for convenience.[9]

Step 1 For given ϑ^* the following minimization problem for $D^{(z)}$ has to be solved for each $z = 1, \ldots, Z$, i.e.

$$D^{(z)} = \arg\min_{D^{(z)}} L_D(\vartheta^*, H(D^{(z)})\vartheta^{(z)}), \quad \text{s.t.} \quad D^{(z)\prime}D^{(z)} = I_K. \tag{3.19}$$

This minimization problem resembles the orthogonal Procrustes (OP) problem, where solutions are discussed and provided by Kristof (1964) and Schönemann (1966), see also Golub and van Loan (2013). The solution involves the following calculations:

1.1 Define $S_z = \bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*$.

[9] The WOP approach as discussed here transforms the output from the unconstrained sampler given a fixed point, i.e. the estimator. In general, the corresponding support of the transformed output does no coincide with the support of the assumed prior distribution. However, this discrepancy is data driven, as the final minimal expected loss estimator is a function of the observed data. In the absence of data, the assumed prior distribution implies the origin as the expected minimal loss estimator and the expected loss minimizing orthogonal transformation, compare Equation (3.14), is undetermined. Thus, the suggested ex-post WOP approach is consistent with the assumed prior distribution.

1.2 Do the singular value decomposition $S_z = U_z M_z V_z'$, where U_z and V_z denote the matrix of eigenvectors of $S_z S_z'$ and $S_z' S_z$, respectively, and M_z denotes a diagonal matrix of singular values, which are the square roots of the eigenvalues of $S_z S_z'$ and $S_z' S_z$. Note that the eigenvalues of $S_z S_z'$ and $S_z' S_z$ are identical.

1.3 Obtain the orthogonal transformation matrix $D^{(z)} = U_z V_z'$.

For further details on the derivation of this solution, see Schönemann (1966). Note that if the dispersion between the cross sections is rather large, the solution may be improved by considering weights, turning the problem to be solved into a weighted orthogonal Procrustes (WOP) problem, see e.g. Lissitz et al. (1976) and Koschat and Swayne (1991). Thus Step 1.1 above is altered into

1.1a Define $S_z = \bar{\Lambda}^{(z)\prime} W \bar{\Lambda}^*$,

where the weighting matrix W has to be diagonal with strictly positive diagonal elements and is initialized as the inverses of the estimated lengths of the loading vectors, i.e.

$$W = Z \left(\sum_{z=1}^{Z} \sqrt{(\bar{\Lambda}^{(z)} \bar{\Lambda}^{(z)\prime}) \odot I_{(S+1)N}} \right)^{-1}. \tag{3.20}$$

Consecutively, we use as weights a function of the number of factors and the determinants of the estimated covariance matrices, which are a measure invariant to orthogonal transformations, i.e. $W = \mathrm{diag}(w_1, \ldots, w_{(S+1)N})$, where

$$w_i = \det \left(\frac{1}{Z} \sum_{z=1}^{Z} (\bar{\lambda}_i^{(z)} - \bar{\lambda}_i^*)' (\bar{\lambda}_i^{(z)} - \bar{\lambda}_i^*) \right)^{-\frac{1}{K}}, \quad i = 1, \ldots, (S+1)N, \tag{3.21}$$

with $\bar{\lambda}_i^{(z)}$ and $\bar{\lambda}_i^*$ denoting the ith row vector of $\bar{\Lambda}^{(z)}$ and $\bar{\Lambda}^*$ respectively. The weighting scheme scales the loadings in such a way that the estimated covariance matrix has determinant 1 for each variable.

Step 2 Choose $\bar{\Lambda}^*$ and Σ^* as implied by $\overline{H(D)\vartheta}$ with $P = 0$.

For arbitrary initial choices of ϑ^* taken from the unconstrained sampler output, less than ten iterations usually suffice to achieve convergence to a fixed point ϑ^* providing the Bayes estimator. Convergence is assumed if the sum of squared deviations between two successive ϑ^* does not exceed a predefined threshold value, where we use 10^{-9}. The iterative procedure of the algorithm suggests to used the transformed output of the unconstrained sample, i.e. $H(D^{(z)})\vartheta^{(z)}$, as input for the next iteration, thus reducing required computer memory capacities.

The following proposition summarizes the suggested ex-post approach for the static factor model.

Proposition 3.3.1. *The ex-post WOP approach solves the rotation problem for the static factor model.*

Proof. The orthogonal matrix $D^{(z)}$ that minimizes the loss function in Equation (3.19) representing the orthogonal Procrustes problem is unique conditional on almost every $\vartheta^{(z)}$ and ϑ^*, where the elements in $\vartheta^{(z)}$ are random variables following a nondegenerate posterior probability distribution as implied by the chosen prior distributions. The availability of a unique solution to the orthogonal Procrustes problem providing a minimum is shown by Kristof (1964), Schönemann (1966) and Golub and van Loan (2013) and for the weighted orthogonal Procrustes problem by Lissitz et al. (1976). Following Golub and van Loan (2013) the minimization problem stated in Equation (3.19) is equivalent to the maximization of $\text{tr}(D^{(z)\prime}\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)$, where the maximizing $D^{(z)}$ can be found by calculation of the singular value decomposition of $\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*$. If $U_z(\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)V_z' = M_z = \text{diag}\left(m_z^{(1)}, \ldots, m_z^{(K)}\right)$ is the singular value decomposition of this matrix and we define the orthogonal matrix $R_z = V_z' D^{(z)\prime} U_z$, then

$$\text{tr}(D^{(z)\prime}\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*) = \text{tr}(D^{(z)\prime}U_z M_z V_z') = \text{tr}(R_z M_z) \leq \sum_{k=1}^{K} m_z^{(k)}.$$

The upper bound is then attained by setting $D^{(z)} = U_z V_z'$, which implies $R_z = I_K$. Note that there exist points, however, where at least one singular value of $\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*$ is zero. In these cases, the left and right eigenvectors related to these singular values are not uniquely determined and thus no unique solution to the orthogonal Procrustes problem exists. However, these points occur with probability zero.

The rotation problem implies that within the parameter space pairs of points can be defined, where the two points are pairwise orthogonal transformations of each other according to Equation (3.5). Denote such a pair as $\vartheta^{(1)}$ and $\vartheta^{(2)}$ with $\vartheta^{(2)} = H(D_0)\vartheta^{(1)}$, where D_0 is an orthogonal matrix. To show that the rotation problem is solved by the suggested ex-post approach, one has to show that no such pairs can be defined after postprocessing. After postprocessing, $\vartheta^{(1)}$ and $\vartheta^{(2)}$ take the form $H(D_1)\vartheta^{(1)}$ and $H(D_2)\vartheta^{(2)}$ respectively, where D_i, $i = 1, 2$ implies minimal loss with regard to ϑ^*. Since D_1 and D_2 are uniquely defined as shown above and $H(D_2)\vartheta^{(2)} = H(D_0 D_2)\vartheta^{(1)}$ we have consequently $D_0 D_2 = D_1$, where we use the fact that the product of two orthogonal matrices is itself an orthogonal matrix, and orthogonal matrices commute. Assuming without loss of generality that $D_1 = I_K$, we have $D_2 = D_0'$. This implies that after postprocessing all points that can be represented as orthogonal transformations of $\vartheta^{(1)}$ are collapsed into $\vartheta^{(1)}$ as the point invoking minimal loss and thus enter the parameter estimation as $\vartheta^{(1)}$. □

Next, we consider the case of the dynamic factor model with $P > 0$. The corresponding ex-post approach is based on an extended loss function considering the dynamic factor structure as well.

Dynamic Factor Model The algorithm for the dynamic factor model differs with regard to Step 1 from the algorithm presented for the static factor model.

Step 1 For given ϑ^* the following minimization problem for $D^{(z)}$ has to be solved for each $z = 1, \ldots, Z$, i.e.

$$D^{(z)} = \arg\min_{D^{(z)}} L_D(\vartheta^*, H(D^{(z)})\vartheta^{(z)}), \quad \text{s.t.} \quad D^{(z)\prime}D^{(z)} = I_K. \tag{3.22}$$

The solution is based on numerical optimization using a parametrization of $D^{(z)}$ ensuring orthogonality. Since every orthogonal matrix D can be decomposed into a reflection matrix B with $\det(B) = \det(D) = \pm 1$ and a corresponding rotation matrix which can be factorized into $\frac{K(K-1)}{2} = |\{(i,j) : i,j \in \{1,\ldots,K\}, j > i\}|$ Givens rotation matrices, we can parameterize any orthogonal matrix as

$$D = \begin{cases} D_+ = B_+ \prod_{(i,j):i,j\in\{1,\ldots,K\},j>i} G_{i,j,K} & \text{if } \det(D) = 1, \\ D_- = B_- \prod_{(i,j):i,j\in\{1,\ldots,K\},j>i} G_{i,j,K} & \text{if } \det(D) = -1, \end{cases} \tag{3.23}$$

where

$$B_+ = \begin{pmatrix} I_{K-1} & 0 \\ 0 & 1 \end{pmatrix}, \quad B_- = \begin{pmatrix} I_{K-1} & 0 \\ 0 & -1 \end{pmatrix} \tag{3.24}$$

and

$$G_{i,j,K} = \begin{pmatrix} g_{1,1} & \cdots & g_{1,K} \\ \vdots & & \vdots \\ g_{K,1} & \cdots & g_{K,K} \end{pmatrix}, \text{ with } g_{r,s} = \begin{cases} 1, & \text{for } i \neq r = s \neq j \\ \cos(\gamma_{(i,j)}), & \text{for } r = s = i \text{ and } r = s = j, \\ -\sin(\gamma_{(i,j)}), & \text{for } r = j, s = i, \\ \sin(\gamma_{(i,j)}), & \text{for } r = i, s = j, \\ 0, & \text{else,} \end{cases} \tag{3.25}$$

and $\gamma_{(i,j)} \in [-\pi, \pi)$ for all $\{(i,j) : i,j \in \{1,\ldots,K\}, j > i\}$.[10] This parametrization allows for a numerical optimization providing two matrices $D_-^{(z)}$ and $D_+^{(z)}$, where $D^{(z)}$ is then chosen as

$$D^{(z)} = \underset{D_-^{(z)}, D_+^{(z)}}{\operatorname{argmin}} \{L_D(\vartheta^*, H(D_-^{(z)})\vartheta^{(z)}), L_D(\vartheta^*, H(D_+^{(z)})\vartheta^{(z)})\}. \tag{3.26}$$

As the starting value for the numerical optimization we choose the solution defined by the WOP algorithm applied to $L_D(\vartheta^*, H(D^{(z)})\vartheta^{(z)})$ with $P = 0$ only. Convergence is quickly achieved and the overall improvement of the target value is generally very small, lying below 3% for all considered data scenarios.[11]

[10] This parametrization resembles the one used by Anderson et al. (1987), which is different with respect to the domain of the angular parameters, which is $\gamma_{(i,j)} \in [-\frac{\pi}{2}, \frac{\pi}{2})$, whereas in our decomposition, it is $\gamma_{(i,j)} \in [-\pi, \pi)$. Extending the domain accordingly allows to reduce the number of reflection parameters from K to 1, hence, our approach is more parsimonious with respect to the number of parameters, and, having all but one of the parameters living in the continuous space, is more easy to handle in optimizations.

[11] The accuracy of the numerical optimization procedure has been assessed via comparison of the numerical with the analytical solution in the static case.

Step 2 Choose $\vartheta^* = \overline{H(D)\vartheta}$.

Following the line of arguments presented in case of the static factor model, we state the properties of the proposed ex-post approach towards the rotation problem for the dynamic factor model in form of two propositions.

Proposition 3.3.2. *The orthogonal matrix $D^{(z)}$ that minimizes the loss function in Equation (3.22) is unique conditional on almost every $\vartheta^{(z)}$ and ϑ^*, where the elements in $\vartheta^{(z)}$ are random variables following a non-degenerate posterior probability distribution as implied by the prior distributions.*

Proof. The proof of Proposition 3.3.2 is given in Appendix 3.B. □

Proposition 3.3.3. *The ex-post WOP approach solves the rotation problem for the dynamic factor model.*

Proof. Given the results from Proposition 3.3.2, the proof is completed using the same line of argument as presented in the proof of Proposition 3.3.1. □

Given that these algorithms provide the minimal expected loss estimators and at convergence a sample $\{H(\tilde{D}^{(z)})\vartheta^{(z)}\}_{z=1}^{Z}$ not subject to orthogonally mixing anymore, it should be highlighted that any minimal expected loss estimator obtained as a transformation of the sample based on a single orthogonal matrix, say D_*, implies the same loss, as follows from the characterization of the minima given in Equation (3.18).[12] Hence we have for an arbitrary orthogonal matrix D_* with $H(D_*)'H(D_*) = I$ and $H(D_*)H(\tilde{D}^{(z)}) = H(D_*\tilde{D}^{(z)})$

$$\begin{aligned}
&\frac{1}{Z}\sum_{z=1}^{Z} \operatorname{tr}\left[\left(H(\tilde{D}^{(z)})\vartheta^{(z)} - \overline{H(\tilde{D})\vartheta}\right)' H(D_*)'H(D_*)\left(H(\tilde{D}^{(z)})\vartheta^{(z)} - \overline{H(\tilde{D})\vartheta}\right)\right] \quad (3.27)\\
&= \frac{1}{Z}\sum_{z=1}^{Z} \operatorname{tr}\left[\left(H(D_*\tilde{D}^{(z)})\vartheta^{(z)} - \overline{H(D_*\tilde{D})\vartheta}\right)' \left(H(D_*\tilde{D}^{(z)})\vartheta^{(z)} - \overline{H(D_*\tilde{D})\vartheta}\right)\right].
\end{aligned}$$

Thus, after approximating the posterior distribution via the output of the Gibbs sampler and applying the WOP approach, D_* can be chosen freely according to interpretational considerations. For example, criteria like Varimax and Quartimax as well as a PLT form can be applied to determine D_*.[13] It is also valid to compare different interpretations based on the same estimation. We consider it as a major advantage that estimation and interpretation of the factor model are clearly separated in the WOP approach. Hence, the WOP approach is purely exploratory and order invariant.

The following section provides a comparison between the suggested ex-post approach and the ex-ante PLT approach.

[12] This property for optimal estimators applies also in the context of finite mixture models with respect to permutations.

[13] Note that the transformation that results in PLT form of the point estimator being applied to all draws does not cause the complete posterior distribution to fulfill the constraints imposed by the ex-ante PLT approach.

3.4 Comparison of the Ex-Post WOP Approach to the Ex-Ante PLT Approach

To illustrate some of the advantages of the suggested ex-post WOP approach, we compare it to the ex-ante PLT approach as suggested by Geweke and Zhou (1996) for Bayesian factor analysis. We especially discuss two issues that are often associated with the ex-ante approach in the literature, namely order dependency and multimodality. The ex-ante PLT approach of Geweke and Zhou (1996) is designed as follows. The parameter space of the loadings is constrained to a positive lower triangular matrix, i.e.

$$\bar{\Lambda} = \begin{pmatrix} \lambda_{1,1} & \lambda_{2,1} & \dots & & & \lambda_{(S+1)N,1} \\ 0 & \lambda_{2,2} & \lambda_{3,2} & \dots & & \lambda_{(S+1)N,2} \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & \lambda_{K,K} & \dots & \lambda_{(S+1)N,K} \end{pmatrix}', \quad \text{with} \quad \lambda_{i,i} > 0, i = 1, \dots, K. \tag{3.28}$$

In maximum likelihood factor analysis this pattern can be obtained by constraining the upper triangular elements of the loadings matrix accordingly. In Bayesian factor analysis, the parameter space can be restricted by using appropriately defined prior distributions. The prior distributions for $\lambda_{i,k}$ with $i < k$ are Dirac Delta distributions and thus fully informative, whereas the prior distributions with regard to $(\lambda_{i,1}, \dots, \lambda_{i,i})'$ for $i \leq K$ take the form of i-variate normal distributions truncated below at zero for $\lambda_{i,i}$.[14] Thus, the ex-ante PLT approach differs from ours in two ways. Firstly, the rotation problem is solved ex ante via priors instead of ex post, and secondly, the PLT approach constrains the parameter space, while the WOP approach postprocesses the output of an unconstrained Gibbs sampler, removing the part of the variation relative to a fixed point that is attributable to orthogonal transformations,

The first of the aforementioned two issues we discuss here is order dependency. This means that inference results depend on the ordering of the variables, or, equivalently, on which elements of Λ the fully informative and truncated priors are placed. This has been observed e.g. by Lopes and West (2004), Carvalho et al. (2008) or recently Chan et al. (2013). The reason for this order dependence can be motivated as follows. Consider the factor model given as $y_t = \sum_{s=0}^{S} \Lambda_s f_{t-s} + e_t$. Then consider an $N \times N$ permutation matrix O that is premultiplied to y_t and relocates at least one factor founder thus resulting in a reordering of the variables given as $Oy_t = O\sum_{s=0}^{S} \Lambda_s f_{t-s} + Oe_t$. When $\bar{\Lambda}$ has PLT form, then $(I_{S+1} \otimes O)\bar{\Lambda}$ almost surely does not have PLT form, since the set of matrices satisfying the PLT constraints under both orderings have probability zero. This implies that almost all admissible points under one set of PLT constraints are inadmissible under a different set of PLT constraints. Consider the

[14] In a similar approach, Aguilar and West (2000) use a degenerate prior distribution, whose probability mass is concentrated at one for the diagonal elements, see e.g. Ishwaran and Rao (2005). This also solves the scaling indeterminacy, so the variances of the factor innovations can be freely estimated. Yet another approach follows the scheme by Jöreskog (1979b), where the top $K \times K$ section of the loadings matrix is constrained to the identity matrix. In turn, all elements of the covariance matrix of the factor innovations can be freely estimated. This approach is discussed in more detail in Bai and Wang (2012).

transformation of a (posterior) distribution that satisfies the PLT constraints under a certain ordering of the data (PLT—O_1) into a distribution that satisfies the PLT constraints under a different ordering of the data (PLT—O_2). The transformation from PLT—O_1 to PLT—O_2 can be achieved by an infinite number of orthogonal matrices.[15] For the (posterior) distribution to remain unchanged except for a single orthogonal transformation, there would have to be a unique orthogonal matrix performing this mapping for every $\bar{\Lambda}$ admissible under the first set of constraints onto a $\bar{\Lambda}$ admissible under the second set of constraints. This highlights why the shape of the posterior distribution depends on the ordering of the data under the ex-ante PLT approach. In contrast, under the ex-post WOP approach the orthogonal mixing in the posterior distribution relative to a fixed point is eliminated. If the fixed point is transformed by an orthogonal matrix the orthogonal mixing in the posterior distributions remains eliminated if all points are transformed by the same matrix.

The second issue is multimodality. Imposing a lower triangularity constraint onto $\bar{\Lambda}$ without additionally demanding positive signs on the diagonal elements ensures *local identification*, see Anderson and Rubin (1956). Thus every reflection of a subset of columns of $\bar{\Lambda}$ yields the same likelihood value. Jennrich (1978) calls this phenomenon *transparent* multimodality. As Loken (2005) shows, introducing nonzero constraints leads to another type of multimodality, sometimes referred to as *genuine* multimodality. Whereas the PLT constraints ensure that the parameter space contains only one global mode, they may induce multiple local modes. Following the notion of Millsap (2001), imposing constraints may result in a likelihood, or in the Bayesian setup in a posterior distribution, which has maxima from different equivalence classes, where an equivalence class corresponds to all points that can be transformed into each other by means of the transformation given in Equation (3.5).

To illustrate the issues of order dependence and of multimodality under the PLT approach, we provide a small example that demonstrates the effect of the constraints under the PLT approach on the shape of the likelihood. In this example we use $S = P = 0$ for simplicity. We start with a set of parameters for a model with $K = 2$ orthogonal static factors having unit variance and $N = 10$ variables, of which the first five are arranged in three different orderings, while the ordering of the remaining five stays identical. This data set is simulated using as parameters

$$\bar{\Lambda} = \begin{pmatrix} 0.100 & -0.200 & 0.500 & 0.600 & 0.100 & 0.174 & -0.153 & -0.470 & 0.186 & -0.577 \\ 0.000 & 0.200 & -0.100 & 0.400 & -0.900 & 0.429 & -0.392 & 0.652 & 0.282 & -0.541 \end{pmatrix}' \tag{3.29}$$

and

$$\Sigma = \mathrm{diag}(0.990, 0.920, 0.740, 0.480, 0.180, 0.786, 0.823, 0.354, 0.886, 0.374). \tag{3.30}$$

[15] All matrices $\bar{\Lambda}$ already satisfying both sets of constraints are transformed by the identity matrix, whereas all matrices whose top $K \times K$ section is identical up to a multiplication with a single scalar are transformed by the same orthogonal matrix. Finally, for those matrices whose top $K \times K$ section is singular, there exists no such orthogonal matrix, see also Chan et al. (2013).

The three orderings of the variables and thus of the rows of $\bar{\Lambda}$ we consider are the one in Equations (3.29) and (3.30), denoted as $Y|O_1$, the second with variable ordering 2,3,1,4,5,6,7,8,9,10, denoted as $Y|O_2$, and the third with variable ordering 5,2,1,3,4,6,7,8,9,10, denoted as $Y|O_3$.

We first obtain the principal components estimate for $\bar{\Lambda}$. At this point, we deviate from Bayesian estimation for illustrative reasons and since we want to exclude the impact of numerical issues arising in MCMC context. The principal component estimate for $\bar{\Lambda}$, denoted as $\bar{\Lambda}^{\text{PC}}$, is afterwards transformed by three orthogonal matrices in order to satisfy the PLT constraints for each of the three orderings $Y|O_1$, $Y|O_2$ and $Y|O_3$ respectively. All three estimates attain the same log likelihood value. Next, we consider all possible orthogonal transformations of $\bar{\Lambda}^{\text{PC}}$ under the three orderings.[16]

A permutation implies a label switching, accordingly the constraints on the loadings of the first factor are exchanged with the constraints on the loadings of the second factor. To illustrate the effect of the constraints, all transformations are afterwards subject to the three initial PLT constraints, i.e. all unconstrained parameters are transformed by the orthogonal matrix, any negative values for the two parameters that are constrained to positivity are set to a small value $\epsilon > 0$ and the loading that is set to zero remains zero.[17]

The exercise shall provide us with an approximation how the constraints affect the likelihood with respect to orthogonal transformations. Figure 3.1 shows the results of the exercise. The solid lines correspond to the transformations by means of rotations and the dashed lines to the transformations by means of a permutation and subsequent rotation. While for $Y|O_1$, the likelihood is almost perfectly flat, hence the constraints have almost no effect at all, for $Y|O_2$, the descent from the global mode is also quite flat in one direction, but considerably steeper in the other direction. For $Y|O_3$, the likelihood declines steeply in both directions. This result corroborates the finding in Carvalho et al. (2008) that inference results vary among different orderings of the variables. Aside from the shape, the permutation and subsequent rotation induces a second mode, which is slightly lower than the first one. This mode is even present under the presumably well-behaved third ordering.

[16] Those orthogonal transformations that are rotations can be expressed by a matrix D_+ with $\det(D_+) = 1$, see Equation (3.23). The orthogonal transformations that involve a label switching between the factors or a reflection about a single axis require an orthogonal matrix D_- with $\det(D_-) = -1$, see Equation (3.23). As can be seen from the decomposition of orthogonal matrices described in Equation (3.23), all orthogonal matrices with dimension 2×2 are expressible as a product of a rotation and a reflection about the second axis. Since this axis reflection can be written as

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

where the first matrix on the right hand side is a permutation of the two factors, and the second is a rotation by the angle $\frac{\pi}{2}$, it is possible to evaluate the effect of all matrices D_+ by rotating $\bar{\Lambda}$, and the effect of all matrices D_- by first exchanging the columns of $\bar{\Lambda}$ and then rotating the result.

[17] Note that if there were no constraints, an orthogonal transformation would leave the likelihood value unchanged, and if there was merely the zero constraint, there likelihood profile on the circle would be bimodal under rotations with two identical modes exactly 180 degrees apart, see Loken (2005), and another two modes would evolve under the label switching and rotation from an identical likelihood profile, but shifted by 90 degrees.

The illustrative exercise based on the principal components analysis makes clear that the impact of the PLT constraints is not just a matter of numerical accuracy or the question whether PLT constraints are imposed ex ante or ex post. Even an implementation of the PLT constraints ex post would affect the shape of the likelihood and thus the shape of the posterior distribution.[18] However, if the PLT constraints are imposed ex ante the resulting shape of the likelihood may in some cases cause numerical problems.

To illustrate the consequences for Bayesian analysis, we estimate the model parameters under all three orderings using the PLT approach. We then repeat the estimation with the unconstrained sampler, and finally post-process the results of the unconstrained sampler using the WOP approach.[19] The first column of Figure 3.2 shows the output of the unconstrained Gibbs sampler with respect to the loadings parameters of variable 8 on both factors, $\bar{\lambda}_8$, under the three orderings, plotted as a Gibbs sequence of length 50,000. The second column shows the according output of the constrained sampler, and the third column shows the unconstrained Gibbs output postprocessed with WOP. The following three columns display the same outputs as bivariate contour plots, with the first 20,000 draws discarded. It can be seen that the posterior density from the unconstrained sampler is invariant with respect to orthogonal transformations, whereas the shape of the posterior density from the constrained sampler depends on the ordering, or the chosen set of PLT constraints. The findings are in line with the likelihood profiles from Figure 3.1, i.e. where the likelihood profiles are flat, the constrained posterior densities are more spread out, where they are peaked, the constrained posterior densities are more concentrated. Conversely, the posterior density obtained from the unconstrained Gibbs output postprocessed with WOP has the same shape under all three orderings of the variables, which resembles that for the PLT constrained sampler output obtained under the third ordering of the variables.

3.5 Simulation Study

To evaluate the properties of the proposed ex-post WOP approach, we perform several simulation experiments, where we use the ex-ante PLT approach as a benchmark. The simulation experiments are designed to analyze the convergence, statistical and numerical properties of both approaches. The considered scenarios have the following features in

[18] The PLT constraints could be imposed ex post based on a loss function defined in correspondence with the one in Equation (3.15). Following Stephens (2000), the corresponding loss function could be formulated as

$$-\sum_{i=1}^{K-1}\sum_{k=i+1}^{K}\log\left[I(\lambda_{i,k}(D)=0)\right]-\sum_{i=1}^{K}\log\left[I(\lambda_{i,i}(D)>0)\right]+\operatorname{tr}\left[(H(D)\vartheta-\vartheta^*)'(H(D)\vartheta-\vartheta^*)\right], \quad (3.31)$$

where $\log(0)$ is defined by its limit value $-\infty$ and $\lambda_{i,k}(D)$ denotes the corresponding element from $H(D)\vartheta$. The MC version of this loss function subject to $D'D = I$ can be minimized by performing a QR decomposition of $\bar{\Lambda}$. However, the resulting estimator and an estimator obtained under a different ordering of the data do not invoke the same loss. Further, even if the constraints are enforced ex post the constraints on the parameter space are formulated ex ante.

[19] The priors are chosen as given in Section (3.2) with hyperparameters $\underline{\Upsilon}_0 = 10I_N$ and $\underline{\alpha}_i = \underline{\beta}_i = 1$, $i = 1, \ldots, N$ and incorporating the truncation as implied by the ex-ante PLT approach.

common. First, at least 20% of the variation in the data is explained by the factors. Second, the loading matrices employed in generating the data sets all satisfy the PLT constraints, so the estimates from the ex-ante PLT approach can be directly compared to the parameter values employed in generating the data.[20] We simulate data sets with $T = 100$ or $T = 500$, respectively, using either $N = 30$ or $N = 100$ variables. Each of these scenarios is paired with $K = 2$ or $K = 4$ stationary factors, which are either static or follow vector autoregressive processes of order $P = 1$ or $P = 4$, respectively. Throughout, we set $S = 0$. In all cases, the number of factors is assumed to be known. The prior distribution is chosen as given in Section 3.2. We consider up to six prior hyperparameter scenarios (I to VI).[21]

To ensure comparability between the estimates of the ex-ante PLT approach and the ex-post WOP approach, we take advantage of the possibility to transform the estimates as indicated by Equation (3.27) based on a single orthogonal matrix, where we use the orthogonal matrix minimizing the sum of quadratic distances between the WOP estimates of the loading parameters and the corresponding values of the loading parameters employed in generating the data. We proceed accordingly with the PLT estimates.[22] Additionally, given that the loading matrices employed in generating the data obey the constraints imposed in the ex-ante PLT approach and these constraints are also employed in the ex-ante estimation, we apply an orthogonal transformation to the ex-post WOP estimates of the loading parameters reconstituting the PLT constraints used in generating the simulated data sets.[23] This allows for calculation of root mean squared errors of the parameter estimators based on the value of the parameter vector employed in repeated simulation of the data.

First, we analyze the convergence properties of the PLT approach and the WOP approach based on the unconstrained sampler. Convergence is checked using the test by Geweke (1991), adjusting for autocorrelation in the draws by means of the heteroscedasticity and autocorrelation-robust covariance estimator by Newey and West (1987). We discard the initial 5,000 draws of the sampler and fix the length of the sample to be kept at 10,000 draws. When no convergence is found for the most recent 10,000 draws, the sequence is extended by another 1,000 draws. The burn-in sequence is not allowed to exceed 100,000 draws, so we do not extend the sequence any further then and assume that it will not converge. Convergence statistics

[20] Note that the diagonal elements of the upper $K \times K$ matrix of $\bar{\Lambda}$ are chosen such that they qualify as *factor founders* in the sense of Carvalho et al. (2008), i.e.

$$\lambda_{i,i} = \sqrt{\frac{1}{(S+1)N - i} \sum_{j=i+1}^{(S+1)N} \lambda_{j,i}^2}.$$

[21] In prior scenario I, the with hyperparameters are set equal to $\underline{\Upsilon}_0 = I_N$ and $\underline{\alpha}_i = \underline{\beta}_i = 1$, $i = 1, \ldots, N$, in prior scenario II we have $\underline{\Upsilon}_0 = I_N$ and $\underline{\alpha}_i = 4$, $\underline{\beta}_i = 0.25$, $i = 1, \ldots, N$, in prior scenario III we have $\underline{\Upsilon}_0 = I_N$, and $\underline{\alpha}_i = 0.25$, $\underline{\beta}_i = 4$, $i = 1, \ldots, N$, in prior scenario IV we have $\underline{\Upsilon}_0 = 100 I_N$, and $\underline{\alpha}_i = \underline{\beta}_i = 1$, $i = 1, \ldots, N$, in prior scenario V we have $\underline{\Upsilon}_0 = 100 I_N$ and $\underline{\alpha}_i = 4$, $\underline{\beta}_i = 0.25$, $i = 1, \ldots, N$, in prior scenario VI we have $\underline{\Upsilon}_0 = 100 I_N$ and $\underline{\alpha}_i = 0.25$, $\underline{\beta}_i = 4$, $i = 1, \ldots, N$.

[22] The required transformation can be obtained via solving the orthogonal Procrustes problem for the $\bar{\Lambda}$ used to generate the data and its estimate from the WOP approach, or from the PLT approach, respectively.

[23] This matrix is unique for every WOP estimate of the loadings matrix whose top $K \times K$ matrix has full rank and can be found via performing a QR decomposition of the WOP estimate of the loadings matrix and flipping the negative column signs, or via the results of Theorem A9.8 from Muirhead (1982).

are evaluated for orthogonally invariant quantities only, i.e. the sum of squared loadings per variable, the idiosyncratic error variances, and the determinants of the persistence matrices of the factors at each lag. The total number of parameters monitored is therefore $2N + P$, and convergence is assumed if the Geweke statistic indicates convergence for 90% of the quantities, where the tests use $\alpha = 0.05$. We simulate 50 different samples for each scenario, which do not all converge for a burn-in limited to 100,000 draws. We therefore report both the number of cases where no convergence is attained and the convergence speed for 25 randomly selected converged sequences for each scenario. Table 3.1 shows the results. The scenarios with $K = 2$ factors do not experience any difficulties with respect to convergence, neither for the PLT approach nor for the WOP approach. Inspecting for the scenarios with $K = 2$ factors 25 randomly chosen sequences that converged after at most 100,000 iterations each, we find the average number of iterations required until convergence ranges between 5,480 and 8,160 with an overall average of approximately 6,500 for the PLT approach, and for the WOP approach, the range is between 5,160 and 6,760, with an overall average of approximately 5,600.

The scenarios with $K = 4$ factors require more iterations to converge and sometimes fail altogether. In particular, the scenario with $P = 4$, $N = 100$ and $T = 100$ stands out. It fails to converge for 4 out of 50 samples under the WOP approach and in 21 out of 50 samples under the PLT approach. Occasional non-convergence can be observed in some other scenarios for the PLT approach, while the WOP approach always converges. Inspecting for the scenarios with $K = 4$ factors 25 randomly chosen sequences that converged after at most 100,000 iterations, we find the average number of iterations required until convergence ranges between 6,600 and 42,160 with an overall average of approximately 13,000 for the PLT approach, and for the WOP approach, it ranges between 5,640 and 26,720, with an overall average of approximately 9,000.[24] Although for model setups with larger number of factors the number of convergence failures increases for both approaches, the simulation results indicate that the relative advantage of the ex-post WOP approach is robust under the alternative prior scenarios, reported in the lower part of Table 3.1.

To highlight the statistical accuracy of estimates under the ex-ante PLT approach and the ex-post WOP approach, Table 3.2 shows the root mean-squared errors (RMSE) for the estimates of the loading parameters when we consider minimization of the quadratic distance between the estimated loadings and the loading parameters employed in simulation. Since the models involve up to 2,000 loadings parameters, we only report the 5%, 50%, and 95% quantiles of these RMSEs. With the single exception of the model with $N = 30$, $T = 500$, $K = 2$ and $P = 4$, all reported quantiles are lower for the WOP approach than for the PLT approach. In several models, particularly such models with static factors, the difference is negligible,

[24] Note, however, that monitoring convergence of orthogonally invariant quantities that are functions of the parameters is not the same as monitoring convergence for the directed parameters in the case of the PLT approach. If convergence is indicated for the orthogonally invariant quantities, estimates for the directed parameters may still perform poorly for the reasons discussed in Section 3.4. Yet, focusing on these quantities is the only feasible approach for comparing convergence behavior of the WOP approach and PLT approach. Convergence of the directed parameters, however, always implies convergence of the orthogonally invariant quantities, hence the results can serve as a lower bound for the actual convergence in the PLT approach.

whereas in other models, particularly those with high lag orders in the VAR process generating the factors, results are in favor of the WOP approach. Judging by the median, the RMSE is often similar and tends to be lower for the WOP approach than for the PLT approach. The upper quantiles, however, reveal that particularly in models with a more complex dynamic factor structure, the WOP approach fares better than the PLT approach. Note that in general, estimates for quantities invariant to orthogonal transformations are very similar for the PLT approach and the WOP approach, as seen for the idiosyncratic variances.[25] This consequently also holds for estimates for the product of factors and loadings, i.e. the systematic part of the model. The documented differences are found to be stable across the different prior scenarios. Further, we compare the statistical accuracy of both approaches by applying an orthogonal transformation to the WOP estimates such that they satisfy the PLT constraints, which hold for the parameters used to generate the simulated data, see Table 3.3. In this situation, the relative advantage is even more pronounced across all considered model setups and prior scenarios.

Eventually, we assess the numerical properties of both approaches, using 25 converged sequences. The directed parameter estimates are again transformed as described before. Table 3.4 shows the numerical standard errors for the loading parameters. For all the models, the numerical standard errors are substantially larger for the PLT approach than for the WOP approach, particularly for models with a $K = 4$ factors and more complex persistence patterns in the factors. Looking at parameters invariant to orthogonal transformations, the verdict is different: Table 3.5 shows very similar results for the median numerical standard error for the PLT approach and the WOP approach, while the right tails reveal substantial differences for some models in favor of the WOP approach. The persistence parameters in the factors, again a set of directed parameters, are evaluated in Table 3.6. Once again, the numerical standard errors are small for the ex-post WOP approach, but large for ex-ante PLT approach. Again, the relative advantages are also present across the considered prior scenarios, see the corresponding lower parts of Tables 3.4 to 3.6.

Table 3.7 reports the average time in seconds required per 1,000 iterations of the sampler in the PLT approach and the WOP approach for all considered models, with the standard deviations given in parentheses. The runtime reported for the WOP approach includes the time required for postprocessing, but not the time for the numerical optimization in the dynamic model.[26] The WOP approach generally requires between 5 and 10% less time than the PLT approach. Unsurprisingly, models with $P = 1$ and $P = 4$ require substantially more time than the models with static factors. The difference between the models with $P = 1$ and $P = 4$ is less pronounced. Apart from that, some of the more highly parameterized models require less runtime than the less highly parameterized models, which is owed to the choice

[25] Results not reported here, but available from the authors upon request.

[26] In the simulation studies, this optimization leads to a negligible reduction of the loss of less than 1% in all models. On average, it requires approximately 30 seconds per 1,000 iterations for the models with $K = 2$ and approximately 200 seconds per 1,000 iterations for the models with $K = 4$, irrespective of the value of P.

of parameters, so incidentally, a model with $K = 4$ can be easier to estimate than a model with $K = 2$ and otherwise identical dimensionality.

Altogether, the simulation study shows that both approaches towards the rotation problem yield very similar inference results for parameters invariant to orthogonal transformations. Modest improvements can be obtained by skipping ex-ante constraints and using the WOP approach instead. Conversely, when inference on directed parameters is concerned, the WOP approach provides much better results than the PLT approach. These results hold for statistical as well as numerical properties. Since convergence is checked based on orthogonally invariant quantities only, the poorer performance of the estimates from the PLT approach is likely due to the non-elliptical shape of the posterior distributions and possible multimodality. These properties are induced by the ex-ante PLT constraints and make the posterior distribution difficult to handle. Using the unconstrained sampler and postprocessing its output by the WOP approach prevents such problems.

3.6 Empirical Example

To further illustrate the WOP approach, we apply it to a data set of $N = 120$ macroeconomic time series taken from Bernanke et al. (2005). The time series are measured at monthly frequency over the period from January 1959 until August 2001, and undergo different types transformations to ensure stationarity. These transformations are described in detail by Bernanke et al. (2005) and also encompass demeaning and standardization. We replicate one of the model setups of Bernanke et al. (2005), with $S = 0$, $K = 4$ factors and $P = 7$ lags in the factor dynamics.[27] In the following, the priors are chosen as given in Chapter 3.2 with hyperparameters $\underline{\Upsilon}_0 = I_N$, and $\underline{\alpha}_i = \underline{\beta}_i = 1$, $i = 1, \ldots, N$, and the truncation and zero constraints additionally imposed in the PLT approach. A Gibbs sequence of $20,000$ iterations is retained after a burn-in of at least $10,000$ iterations, which is extended until convergence is attained according to Geweke's criterion.

To highlight the numerical advantages of the WOP approach, we perform 20 repeated estimations of a specific ordering of the data using the PLT approach and the WOP approach. The four data series that are then affected by the PLT constraints, hence the factor founders, are the federal funds rate (FYFF), the industrial production (IP), the monetary base (FM2), and the NAPM (National Association of Purchasing Management) commodity price index (PMCP). To make results under both approaches comparable, the posterior means under WOP are rotated such that they satisfy the PLT constraints.[28] Figure 3.3 shows the 20 estimates for all four factors under both approaches. The correlation between the federal funds rate and the first factor is 0.9989 for the PLT approach and 0.9987 for the WOP

[27] However, instead of considering the Fed Funds rate as an observable factor and the three remaining factors as latent, we assume that all factors are latent. The WOP approach allows for the estimation of factor-augmented vector-autoregressive models as well, however, this is beyond the scope of this paper.

[28] The required orthogonal matrix is found as in the simulation study by first performing the QR decomposition of the estimated loadings matrix under the WOP approach, see e.g. Golub and van Loan (2013). The remaining model parameters can then be transformed accordingly.

approach. The numerical variation of the first two factors is slightly larger under the PLT approach compared to the WOP approach, while it is much larger for the last two factors.

Next, to show the order invariance of the WOP approach, we consider 20 different orderings of the time series and estimate the accordingly specified factor model for each of them both by the PLT approach and the WOP approach. Estimation is based here on a single set of common random numbers to mitigate their effect on the numerical precision and thus overall precision of the estimates.[29] Table 3.8 shows the average standard deviation over all 480 loadings parameters for each of the considered 20 orderings of the variables. While the average standard deviations under the WOP approach vary only little, they deviate substantially from each other under the PLT approach. Moreover, the smallest average standard deviation under the PLT approach is almost as small as under the WOP approach, while all other average standard deviations under the PLT approach are bigger. Figure 3.4 shows the results of the factor estimates, where the results under both approaches are orthogonally transformed such that the PLT structure holds with respect to the initial four factor founders. The results under the WOP approach are virtually identical compared to the first exercise, where the average correlation between the first factor and FYFF is again 0.9987. This illustrates that the estimation under the WOP approach is indeed invariant to the ordering of the variables. Results obtained under the PLT approach show clear variations, which are much bigger than under the conveniently ordered time series. The average correlation between the first factor and FYFF is now only 0.8418, with 12 out of the 20 orderings reaching a correlation of more than 0.99, but 6 out of them failing to exceed even 0.7. Orthogonally mapping pairs of parameter estimates obtained under the different orderings onto each other yields an average deviation, measured as the Frobenius norm of the matrix of differences, that is about 15 times larger for the estimates obtained from the PLT approach, compared to those obtained from the WOP approach. This underlines the order dependency of the estimation under the PLT approach. It must be noted that while rather convenient orderings for the PLT approach exist, they still do not outperform the results obtained under the WOP approach.

In summary, the results of the empirical example underline the results of the simulation study and highlight that the WOP approach has favorable numerical properties and provides order invariant inference.

3.7 Conclusion

We propose an ex-post approach to solve the rotation problem in Bayesian analysis of static and dynamic factor models. The PLT approach is commonly used and imposes constraints on the loadings matrix ex ante by using truncated and degenerate prior distributions on its upper triangular elements. Inference results based on the PLT approach have been observed to be order dependent. Thus, we suggest to refrain from imposing ex-ante constraints via

[29] The number of different orderings, or choice of factor founders, is prohibitively large, attaining 197 million, so we choose only a small random sample.

according prior distributions. Instead, we propose to use an orthogonally unconstrained sampler, which does not introduce constraints by the according prior distributions, but instead is based on prior distributions for all model parameters that are invariant under orthogonal transformations. Using the orthogonally unconstrained sampler also avoids numerical problems that may occur when sampling from truncated distributions. The rotation problem is subsequently solved in a postprocessing step, where the distance between each draw from the unconstrained sampler and a fixed point is minimized. For static models the minimization problem for each draw of the sampler resembles the weighted orthogonal Procrustes (WOP) problem, which has a unique analytic solution except for probability zero events. For dynamic models a unique solution exists as well except for probability zero events, which can be found by using a numerical optimization routine.

The WOP approach has several desirable properties. The shape of the posterior distribution does not depend on the ordering of the data, hence the inference results are likewise not order dependent. Furthermore, estimation and interpretation can be treated separately, as arbitrary rotation procedures, such like Varimax, can be applied to the posterior mean of the postprocessed Gibbs output. In a simulation study as well as in an application to a large macroeconomic data set, we compare the WOP approach with the commonly used PLT. Both exercises confirm the order independence of the WOP approach, which also converges faster and yields lower MC errors.

Acknowledgements

For very helpful comments and thoughtful suggestions they provided on earlier versions of the paper, we thank Sylvia Frühwirth-Schnatter, Uwe Jensen, Sylvia Kaufmann, Roman Liesenfeld, Milan Stehlik, Helga Wagner, the participants of the Statistische Woche 2012, of the European Seminar on Bayesian Econometrics 2012, of the Vienna Workshop on High-Dimensional Time Series 2013, and of the ESEM 2013, and two anonymous referees.

Tables

Table 3.1: Number of sequences not converged after 100,000 iterations (nc) and average length of burn-in for 25 (ab) randomly chosen converged sequences per model.

N	T	P	prior scenario	PLT				WOP			
				$K=2$		$K=4$		$K=2$		$K=4$	
				nc	ab	nc	ab	nc	ab	nc	ab
30	100	0	I	0	6160 (1028)	0	6600 (2041)	0	5280 (614)	0	5640 (1186)
30	100	1	I	0	5840 (1344)	0	11160 (6176)	0	5800 (1384)	0	5760 (1200)
30	100	4	I	0	6760 (3666)	0	13440 (13961)	0	5800 (1414)	0	17120 (13581)
30	500	0	I	0	5480 (714)	0	6680 (1973)	0	5320 (690)	0	5840 (1214)
30	500	1	I	0	6200 (3000)	0	8440 (2022)	0	5360 (638)	0	7320 (2155)
30	500	4	I	0	6160 (2544)	2	14200 (14428)	0	6760 (1665)	0	7240 (2891)
100	100	0	I	0	6000 (1414)	0	5640 (810)	0	5400 (645)	0	6160 (2014)
100	100	1	I	0	6200 (1633)	1	19320 (18538)	0	5560 (1044)	0	7440 (2501)
100	100	4	I	0	7480 (4094)	21	42160 (24535)	0	5800 (1291)	4	26720 (19661)
100	500	0	I	0	6240 (1234)	0	7920 (3651)	0	5160 (473)	0	6760 (2314)
100	500	1	I	0	8160 (2075)	0	8800 (4041)	0	5280 (614)	0	5760 (1012)
100	500	4	I	0	7080 (2971)	1	14280 (8299)	0	5640 (757)	0	7160 (2495)
				prior sensitivity							
30	100	0	II	0	5076 (341)	–	–	0	5060 (245)	–	–
30	100	0	III	0	5168 (469)	–	–	0	5164 (741)	–	–
30	100	0	IV	0	5296 (721)	–	–	0	5300 (912)	–	–
30	100	0	V	0	5292 (737)	–	–	0	5004 (20)	–	–
30	100	0	VI	0	5616 (1345)	–	–	0	5496 (1272)	–	–
30	100	1	II	–	–	13	27175 (22526)	–	–	1	6168 (1964)
30	100	1	III	–	–	11	35622 (21070)	–	–	0	8192 (4979)
30	100	1	IV	–	–	9	28233 (23118)	–	–	3	16540 (18375)
30	100	1	V	–	–	6	33262 (23840)	–	–	1	11200 (13141)
30	100	1	VI	–	–	9	38508 (28077)	–	–	2	32584 (26337)
30	500	4	II	2	8132 (2106)	–	–	0	5836 (1587)	–	–
30	500	4	III	1	8600 (2799)	–	–	0	5256 (705)	–	–
30	500	4	IV	0	8164 (2047)	–	–	0	5252 (695)	–	–
30	500	4	V	0	8164 (2048)	–	–	0	5248 (706)	–	–
30	500	4	VI	0	8436 (2031)	–	–	0	5568 (1601)	–	–

Notes: Heteroskedasticity and autocorrelation consistent (HAC) standard errors in parentheses. Minimum burn-in for each model is 5,000 iterations. Convergence is monitored for orthogonally invariant statistics of the parameters and assumed to hold if Geweke's (1991) test does not reject the Null hypothesis of convergence for at least 90% of the parameters with $\alpha = 5\%$.

Table 3.2: Distribution quantiles of the RMSE across the loading parameters from 25 randomly chosen converged sequences per model.

N	T	K	P	prior scenario	PLT			WOP		
					q_{05}	q_{50}	q_{95}	q_{05}	q_{50}	q_{95}
30	100	2	0	I	0.0674	0.1232	0.1814	0.0658	0.1137	0.1612
30	100	2	1	I	0.1851	0.2779	0.4021	0.0583	0.1002	0.1569
30	100	2	4	I	0.2246	0.3770	0.5283	0.1104	0.2428	0.4748
30	100	4	0	I	0.1023	0.2366	0.3625	0.0871	0.1304	0.1851
30	100	4	1	I	0.2449	0.4939	0.7308	0.0752	0.1548	0.2554
30	100	4	4	I	0.3225	0.6588	1.0892	0.1051	0.3103	0.5060
30	500	2	0	I	0.0318	0.0597	0.0835	0.0317	0.0586	0.0796
30	500	2	1	I	0.0521	0.1277	0.2130	0.0403	0.1188	0.1743
30	500	2	4	I	0.1013	0.2242	0.4940	0.0572	0.2135	0.5385
30	500	4	0	I	0.0542	0.1094	0.1907	0.0456	0.0679	0.0941
30	500	4	1	I	0.0865	0.1635	0.2779	0.0463	0.1101	0.2432
30	500	4	4	I	0.1850	0.5224	1.0558	0.0549	0.1881	0.3656
100	100	2	0	I	0.0778	0.1220	0.1734	0.0756	0.1135	0.1616
100	100	2	1	I	0.0968	0.1760	0.2644	0.0616	0.0961	0.1574
100	100	2	4	I	0.1841	0.3244	0.4115	0.0763	0.1269	0.1841
100	100	4	0	I	0.1146	0.2122	0.3265	0.0876	0.1297	0.1831
100	100	4	1	I	0.2452	0.4462	0.7669	0.0783	0.1216	0.1801
100	100	4	4	I	0.4041	0.6458	0.9625	0.1349	0.4840	1.0730
100	500	2	0	I	0.0439	0.0633	0.0871	0.0387	0.0541	0.0729
100	500	2	1	I	0.0350	0.0565	0.0779	0.0302	0.0455	0.0720
100	500	2	4	I	0.0638	0.0871	0.1193	0.0423	0.0671	0.1334
100	500	4	0	I	0.0647	0.1059	0.2109	0.0451	0.0656	0.0899
100	500	4	1	I	0.0590	0.0903	0.1477	0.0479	0.0757	0.1608
100	500	4	4	I	0.4626	0.8145	1.4613	0.1417	0.3435	0.7268
					prior sensitivity					
30	100	2	0	II	0.0796	0.1200	0.1670	0.0793	0.1182	0.1651
30	100	2	0	III	0.0793	0.1197	0.1665	0.0790	0.1165	0.1648
30	100	2	0	IV	0.1849	0.2749	0.3934	0.1956	0.2901	0.4189
30	100	2	0	V	0.1842	0.2763	0.3909	0.1965	0.2917	0.4203
30	100	2	0	VI	0.1855	0.2728	0.3949	0.1991	0.2975	0.4266
30	100	4	1	II	0.1682	0.2725	0.4349	0.1024	0.1468	0.2271
30	100	4	1	III	0.1090	0.2090	0.3291	0.1025	0.1453	0.2322
30	100	4	1	IV	0.1957	0.3945	0.8430	0.2048	0.4157	0.8629
30	100	4	1	V	0.1855	0.3838	0.8478	0.2071	0.4178	0.8693
30	100	4	1	VI	0.1826	0.3830	0.8293	0.2060	0.4196	0.9055
30	500	2	4	II	0.1555	0.2979	0.4553	0.0666	0.2072	0.5335
30	500	2	4	III	0.1561	0.2984	0.4556	0.0672	0.2081	0.5347
30	500	2	4	IV	0.2155	0.3906	0.5801	0.1095	0.2924	0.6789
30	500	2	4	V	0.2156	0.3915	0.5809	0.1090	0.2923	0.6782
30	500	2	4	VI	0.2147	0.3890	0.5783	0.1096	0.2924	0.6789

Notes: Involved estimators are orthogonally transformed, such that the distance between the estimated and simulated parameters is minimized.

Table 3.3: Distribution quantiles of the RMSE across the loading parameters from 25 randomly chosen converged sequences per model.

N	T	K	P	prior scenario	PLT			WOP		
					q_{05}	q_{50}	q_{95}	q_{05}	q_{50}	q_{95}
					minimized quadratic distance between estimated and data generating loadings					
30	100	2	0	I	0.0674	0.1232	0.1814	0.0658	0.1137	0.1612
30	100	4	1	I	0.2449	0.4939	0.7308	0.0752	0.1548	0.2554
100	500	2	4	I	0.0638	0.0871	0.1193	0.0423	0.0671	0.1334
30	100	2	0	II	0.0798	0.1319	0.2031	0.0782	0.1171	0.1596
30	100	4	1	II	0.2666	0.4906	0.7305	0.0990	0.1489	0.2380
30	500	2	4	II	0.0811	0.2495	0.4820	0.0695	0.2122	0.5379
30	100	2	0	III	0.0783	0.1310	0.2059	0.0796	0.1157	0.1594
30	100	4	1	III	0.2484	0.4225	0.6207	0.1005	0.1483	0.2398
30	500	2	4	III	0.0865	0.2515	0.4842	0.0698	0.2124	0.5370
30	100	2	0	IV	0.1824	0.2603	0.3653	0.2005	0.2852	0.4183
30	100	4	1	IV	0.1883	0.3445	0.7317	0.2028	0.4092	0.8996
30	500	2	4	IV	0.1458	0.3365	0.5883	0.1124	0.2917	0.6771
30	100	2	0	V	0.1818	0.2588	0.3619	0.2008	0.2867	0.4169
30	100	4	1	V	0.2396	0.4085	0.6859	0.2001	0.4160	0.8714
30	500	2	4	V	0.1445	0.3358	0.5896	0.1123	0.2915	0.6758
30	100	2	0	VI	0.1819	0.2611	0.3666	0.2027	0.2944	0.4273
30	100	4	1	VI	0.1778	0.3294	0.6347	0.2060	0.4236	0.8774
30	500	2	4	VI	0.1473	0.3371	0.5870	0.1126	0.2926	0.6778
					PLT form of estimates					
30	100	2	0	I	0.1298	0.2038	0.2937	0.1311	0.1968	0.2777
30	100	4	1	I	0.4898	0.9023	1.5676	0.2266	0.4353	0.8738
30	500	2	4	I	0.2726	0.4438	0.7467	0.0898	0.2871	0.5828
30	100	2	0	II	0.1320	0.2076	0.2993	0.1307	0.1968	0.2772
30	100	4	1	II	0.4936	0.8353	1.5783	0.2203	0.4229	0.7548
30	500	2	4	II	0.0892	0.3467	0.7687	0.0903	0.2876	0.5848
30	100	2	0	III	0.1317	0.2119	0.3062	0.1303	0.1964	0.2748
30	100	4	1	III	0.5360	0.8642	1.3960	0.2250	0.4301	0.8680
30	500	2	4	III	0.2102	0.4056	0.7659	0.0893	0.2872	0.5830
30	100	2	0	IV	0.2070	0.2948	0.4777	0.2264	0.3274	0.4907
30	100	4	1	IV	0.4756	0.9584	1.6396	0.3718	0.6792	1.2025
30	500	2	4	IV	0.1343	0.3357	0.8433	0.0856	0.3654	0.7123
30	100	2	0	V	0.2086	0.2932	0.4794	0.2281	0.3269	0.4912
30	100	4	1	V	0.5217	0.9991	1.8030	0.3737	0.6550	1.1643
30	500	2	4	V	0.1332	0.3416	0.8509	0.0857	0.3654	0.7108
30	100	2	0	VI	0.2080	0.2962	0.4769	0.2331	0.3318	0.5002
30	100	4	1	VI	0.5139	1.0201	1.8077	0.3783	0.6786	1.1891
30	500	2	4	VI	0.1373	0.3328	0.8367	0.0854	0.3664	0.7127

Table 3.4: Distribution quantiles of the average MC error across the loading parameters from 25 randomly chosen converged sequences.

N	T	K	P	prior scenario	PLT			WOP		
					q_{05}	q_{50}	q_{95}	q_{05}	q_{50}	q_{95}
30	100	2	0	I	0.0072	0.0168	0.0364	0.0021	0.0057	0.0096
30	100	2	1	I	0.0364	0.1196	0.3328	0.0029	0.0058	0.0101
30	100	2	4	I	0.0095	0.0238	0.0372	0.0028	0.0059	0.0096
30	100	4	0	I	0.0377	0.0814	0.1557	0.0014	0.0053	0.0085
30	100	4	1	I	0.0476	0.1743	0.3184	0.0038	0.0091	0.0199
30	100	4	4	I	0.0622	0.1090	0.2060	0.0031	0.0082	0.0140
30	500	2	0	I	0.0044	0.0093	0.0153	0.0011	0.0024	0.0037
30	500	2	1	I	0.0054	0.0119	0.0270	0.0015	0.0036	0.0052
30	500	2	4	I	0.0081	0.0166	0.0246	0.0019	0.0051	0.0097
30	500	4	0	I	0.0189	0.0586	0.1335	0.0010	0.0033	0.0045
30	500	4	1	I	0.0339	0.1512	0.3231	0.0013	0.0048	0.0087
30	500	4	4	I	0.0451	0.1643	0.3405	0.0015	0.0046	0.0066
100	100	2	0	I	0.0289	0.0550	0.0939	0.0026	0.0040	0.0059
100	100	2	1	I	0.0145	0.0712	0.1271	0.0039	0.0061	0.0085
100	100	2	4	I	0.0473	0.1249	0.2145	0.0033	0.0060	0.0110
100	100	4	0	I	0.0307	0.0555	0.0917	0.0035	0.0055	0.0082
100	100	4	1	I	0.1009	0.1891	0.3768	0.0045	0.0071	0.0099
100	100	4	4	I	0.0560	0.1255	0.2894	0.0108	0.0202	0.0375
100	500	2	0	I	0.0054	0.0263	0.0488	0.0027	0.0036	0.0050
100	500	2	1	I	0.0131	0.0286	0.0430	0.0030	0.0048	0.0062
100	500	2	4	I	0.0212	0.0590	0.0806	0.0030	0.0047	0.0074
100	500	4	0	I	0.0308	0.0581	0.1637	0.0032	0.0045	0.0059
100	500	4	1	I	0.0297	0.0588	0.1719	0.0035	0.0051	0.0069
100	500	4	4	I	0.1023	0.2556	0.5931	0.0037	0.0074	0.0150
					prior sensitivity					
30	100	2	0	II	0.0032	0.0050	0.0084	0.0030	0.0043	0.0062
30	100	2	0	III	0.0037	0.0053	0.0081	0.0028	0.0044	0.0075
30	100	2	0	IV	0.0098	0.0153	0.0230	0.0072	0.0122	0.0202
30	100	2	0	V	0.0091	0.0128	0.0183	0.0079	0.0114	0.0172
30	100	2	0	VI	0.0083	0.0126	0.0209	0.0085	0.0141	0.0243
30	100	4	1	II	0.1009	0.1985	0.3167	0.0035	0.0068	0.0178
30	100	4	1	III	0.1380	0.2338	0.3416	0.0039	0.0079	0.0189
30	100	4	1	IV	0.1205	0.2634	0.5026	0.0100	0.0179	0.0567
30	100	4	1	V	0.1347	0.2752	0.4627	0.0107	0.0187	0.0550
30	100	4	1	VI	0.0670	0.1513	0.2462	0.0110	0.0214	0.0491
30	500	2	4	II	0.0798	0.2205	0.5700	0.0036	0.0056	0.0094
30	500	2	4	III	0.0531	0.1415	0.2371	0.0033	0.0058	0.0104
30	500	2	4	IV	0.0360	0.1406	0.2463	0.0057	0.0083	0.0139
30	500	2	4	V	0.1297	0.2639	0.4675	0.0039	0.0069	0.0140
30	500	2	4	VI	0.1204	0.2824	0.6267	0.0046	0.0066	0.0115

Table 3.5: Distribution quantiles of the average MC error across the idiosyncratic variances from 25 randomly chosen converged sequences.

N	T	K	P	prior scenario	PLT			WOP		
					q_{05}	q_{50}	q_{95}	q_{05}	q_{50}	q_{95}
30	100	2	0	I	0.0007	0.0012	0.0025	0.0007	0.0015	0.0028
30	100	2	1	I	0.0007	0.0013	0.0029	0.0006	0.0013	0.0029
30	100	2	4	I	0.0006	0.0013	0.0028	0.0004	0.0014	0.0030
30	100	4	0	I	0.0010	0.0019	0.0114	0.0009	0.0019	0.0030
30	100	4	1	I	0.0011	0.0021	0.3164	0.0011	0.0018	0.0030
30	100	4	4	I	0.0009	0.0021	0.0533	0.0010	0.0018	0.0037
30	500	2	0	I	0.0002	0.0005	0.0013	0.0003	0.0006	0.0012
30	500	2	1	I	0.0002	0.0006	0.0016	0.0003	0.0006	0.0011
30	500	2	4	I	0.0003	0.0006	0.0013	0.0003	0.0006	0.0012
30	500	4	0	I	0.0004	0.0007	0.0031	0.0004	0.0008	0.0015
30	500	4	1	I	0.0004	0.0007	0.0174	0.0004	0.0008	0.0015
30	500	4	4	I	0.0004	0.0007	0.0243	0.0004	0.0008	0.0016
100	100	2	0	I	0.0005	0.0012	0.0028	0.0006	0.0015	0.0031
100	100	2	1	I	0.0006	0.0011	0.0024	0.0006	0.0014	0.0029
100	100	2	4	I	0.0006	0.0014	0.0029	0.0007	0.0014	0.0030
100	100	4	0	I	0.0007	0.0013	0.0032	0.0008	0.0015	0.0034
100	100	4	1	I	0.0007	0.0014	0.0031	0.0008	0.0015	0.0032
100	100	4	4	I	0.0007	0.0015	0.0055	0.0008	0.0016	0.0031
100	500	2	0	I	0.0002	0.0005	0.0011	0.0003	0.0006	0.0012
100	500	2	1	I	0.0002	0.0005	0.0012	0.0003	0.0006	0.0013
100	500	2	4	I	0.0003	0.0005	0.0011	0.0003	0.0006	0.0013
100	500	4	0	I	0.0003	0.0006	0.0012	0.0003	0.0007	0.0013
100	500	4	1	I	0.0003	0.0005	0.0013	0.0004	0.0007	0.0012
100	500	4	4	I	0.0006	0.0027	0.0278	0.0003	0.0008	0.0017
					prior sensitivity					
30	100	2	0	II	0.0008	0.0015	0.0025	0.0008	0.0015	0.0029
30	100	2	0	III	0.0007	0.0012	0.0028	0.0007	0.0011	0.0023
30	100	2	0	IV	0.0008	0.0014	0.0026	0.0009	0.0014	0.0026
30	100	2	0	V	0.0008	0.0015	0.0029	0.0008	0.0014	0.0023
30	100	2	0	VI	0.0006	0.0012	0.0022	0.0008	0.0014	0.0024
30	100	4	1	II	0.0010	0.0020	0.0037	0.0011	0.0020	0.0030
30	100	4	1	III	0.0013	0.0020	0.0043	0.0010	0.0017	0.0029
30	100	4	1	IV	0.0010	0.0024	0.0090	0.0006	0.0016	0.0031
30	100	4	1	V	0.0015	0.0034	0.0079	0.0013	0.0016	0.0031
30	100	4	1	VI	0.0012	0.0025	0.0058	0.0009	0.0016	0.0030
30	500	2	4	II	0.0004	0.0013	0.0064	0.0003	0.0006	0.0011
30	500	2	4	III	0.0004	0.0010	0.0081	0.0004	0.0006	0.0012
30	500	2	4	IV	0.0005	0.0009	0.0062	0.0003	0.0007	0.0012
30	500	2	4	V	0.0004	0.0015	0.0086	0.0004	0.0006	0.0013
30	500	2	4	VI	0.0004	0.0013	0.0075	0.0004	0.0006	0.0011

Table 3.6: Distribution quantiles of the average MC error across the persistence parameters in the factors from 25 randomly chosen converged sequences.

N	T	K	P	prior scenario	PLT			WOP		
					q_{05}	q_{50}	q_{95}	q_{05}	q_{50}	q_{95}
30	100	2	1	I	0.0487	0.0601	0.0701	0.0005	0.0010	0.0027
30	100	2	4	I	0.0023	0.0035	0.0109	0.0013	0.0017	0.0023
30	100	4	1	I	0.0187	0.0363	0.0460	0.0009	0.0012	0.0020
30	100	4	4	I	0.0145	0.0226	0.0454	0.0014	0.0020	0.0029
30	500	2	1	I	0.0021	0.0048	0.0059	0.0006	0.0007	0.0010
30	500	2	4	I	0.0005	0.0021	0.0036	0.0006	0.0008	0.0012
30	500	4	1	I	0.0090	0.0342	0.0941	0.0004	0.0006	0.0011
30	500	4	4	I	0.0070	0.0168	0.0377	0.0004	0.0006	0.0009
100	100	2	1	I	0.0053	0.0141	0.0174	0.0007	0.0011	0.0019
100	100	2	4	I	0.0234	0.0576	0.1044	0.0010	0.0013	0.0027
100	100	4	1	I	0.0252	0.0402	0.0934	0.0006	0.0013	0.0020
100	100	4	4	I	0.0108	0.0191	0.0339	0.0015	0.0029	0.0053
100	500	2	1	I	0.0082	0.0130	0.0174	0.0003	0.0006	0.0013
100	500	2	4	I	0.0015	0.0051	0.0119	0.0004	0.0007	0.0009
100	500	4	1	I	0.0021	0.0072	0.0202	0.0004	0.0007	0.0009
100	500	4	4	I	0.0151	0.0316	0.0959	0.0005	0.0011	0.0035
					prior sensitivity					
30	100	4	1	II	0.0179	0.0414	0.0556	0.0006	0.0010	0.0016
30	100	4	1	III	0.0177	0.0457	0.0973	0.0006	0.0011	0.0017
30	100	4	1	IV	0.0115	0.0293	0.0923	0.0009	0.0019	0.0027
30	100	4	1	V	0.0125	0.0325	0.0872	0.0008	0.0022	0.0033
30	100	4	1	VI	0.0085	0.0174	0.0481	0.0010	0.0023	0.0032
30	500	2	4	II	0.0024	0.0213	0.1839	0.0003	0.0007	0.0012
30	500	2	4	III	0.0019	0.0068	0.0202	0.0006	0.0007	0.0011
30	500	2	4	IV	0.0025	0.0068	0.0175	0.0005	0.0008	0.0019
30	500	2	4	V	0.0035	0.0228	0.1344	0.0006	0.0008	0.0012
30	500	2	4	VI	0.0035	0.0241	0.1846	0.0005	0.0008	0.0016

Table 3.7: Time in seconds elapsed per 1,000 iterations for each model.

N	T	P	PLT		WOP	
			$K = 2$	$K = 4$	$K = 2$	$K = 4$
30	100	0	11.0354 (1.3243)	10.8748 (1.0024)	9.9237 (1.0469)	9.8186 (1.1318)
30	100	1	36.6941 (5.3846)	44.0471 (8.6127)	34.4625 (6.1184)	40.2625 (7.3427)
30	100	4	40.2053 (5.7511)	42.1311 (7.2455)	37.6840 (6.0755)	40.1622 (7.2919)
30	500	0	27.2992 (3.2819)	44.3834 (7.6330)	23.4880 (3.1661)	38.9550 (6.8294)
30	500	1	242.3423 (75.3879)	173.4725 (32.4472)	237.2109 (70.7870)	152.8885 (31.4506)
30	500	4	174.8800 (22.9406)	203.1825 (23.3367)	160.2842 (24.0117)	181.3410 (23.4687)
100	100	0	25.9385 (2.2509)	34.1571 (5.3270)	24.8876 (2.2868)	32.3892 (5.3041)
100	100	1	179.0496 (44.6298)	151.3143 (23.6145)	171.0725 (46.3002)	145.0975 (26.0588)
100	100	4	208.3132 (65.7762)	195.5125 (24.9916)	200.0010 (65.3354)	188.6807 (27.7432)
100	500	0	56.4646 (5.6858)	61.0417 (5.9159)	51.7129 (6.0889)	55.5395 (7.5404)
100	500	1	659.2169 (121.4037)	757.4466 (135.7986)	627.6876 (122.2412)	726.9703 (141.3559)
100	500	4	745.7686 (81.1917)	1399.8993 (209.0987)	719.2335 (94.9956)	1333.1730 (222.4022)

Notes: Postprocessing time is included for the WOP approach. Calculations were performed on intel® i7-4670 (Haswell) processors.

Table 3.8: Average of the 480 posterior standard deviations of loading parameters for 20 different randomly chosen orderings. Corresponding standard deviations are given in parentheses.

ordering	PLT	WOP
1	0.2114 (0.1350)	0.0343 (0.0101)
2	0.0978 (0.0973)	0.0350 (0.0110)
3	0.1394 (0.1410)	0.0347 (0.0103)
4	0.1260 (0.1384)	0.0368 (0.0109)
5	0.1384 (0.1364)	0.0373 (0.0113)
6	0.0349 (0.0105)	0.0344 (0.0108)
7	0.1351 (0.1359)	0.0336 (0.0105)
8	0.0410 (0.0129)	0.0341 (0.0105)
9	0.1400 (0.0942)	0.0340 (0.0108)
10	0.1147 (0.0871)	0.0362 (0.0111)
11	0.1572 (0.1432)	0.0366 (0.0111)
12	0.0479 (0.0165)	0.0349 (0.0103)
13	0.1426 (0.1447)	0.0348 (0.0108)
14	0.1662 (0.1312)	0.0343 (0.0106)
15	0.1559 (0.1501)	0.0376 (0.0095)
16	0.1535 (0.1663)	0.0352 (0.0107)
17	0.1955 (0.1027)	0.0372 (0.0113)
18	0.1368 (0.1433)	0.0343 (0.0104)
19	0.0409 (0.0195)	0.0351 (0.0111)
20	0.1324 (0.1125)	0.0352 (0.0108)

Figures

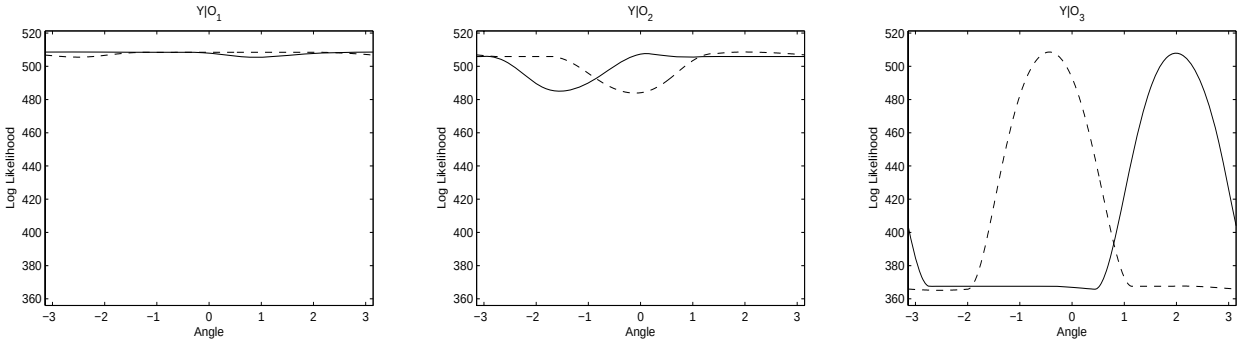

Figure 3.1: Log likelihood values of the principal component estimates, rotated along the circle, with constraints imposed.

Notes: y-axis: log likelihood. x-axis: angle γ of a Givens rotation, where the rotation matrix D_+ is parameterized as $D_+ = ((\cos(\gamma), \sin(\gamma))', (-\sin(\gamma), \cos(\gamma))')$, $\gamma \in (-\pi, \pi)$. The straight line represents the likelihood values that are obtained by first rotating the factors along the circle and then imposing PLT constraints. The dashed line refers to the same exercise but includes a permutation.

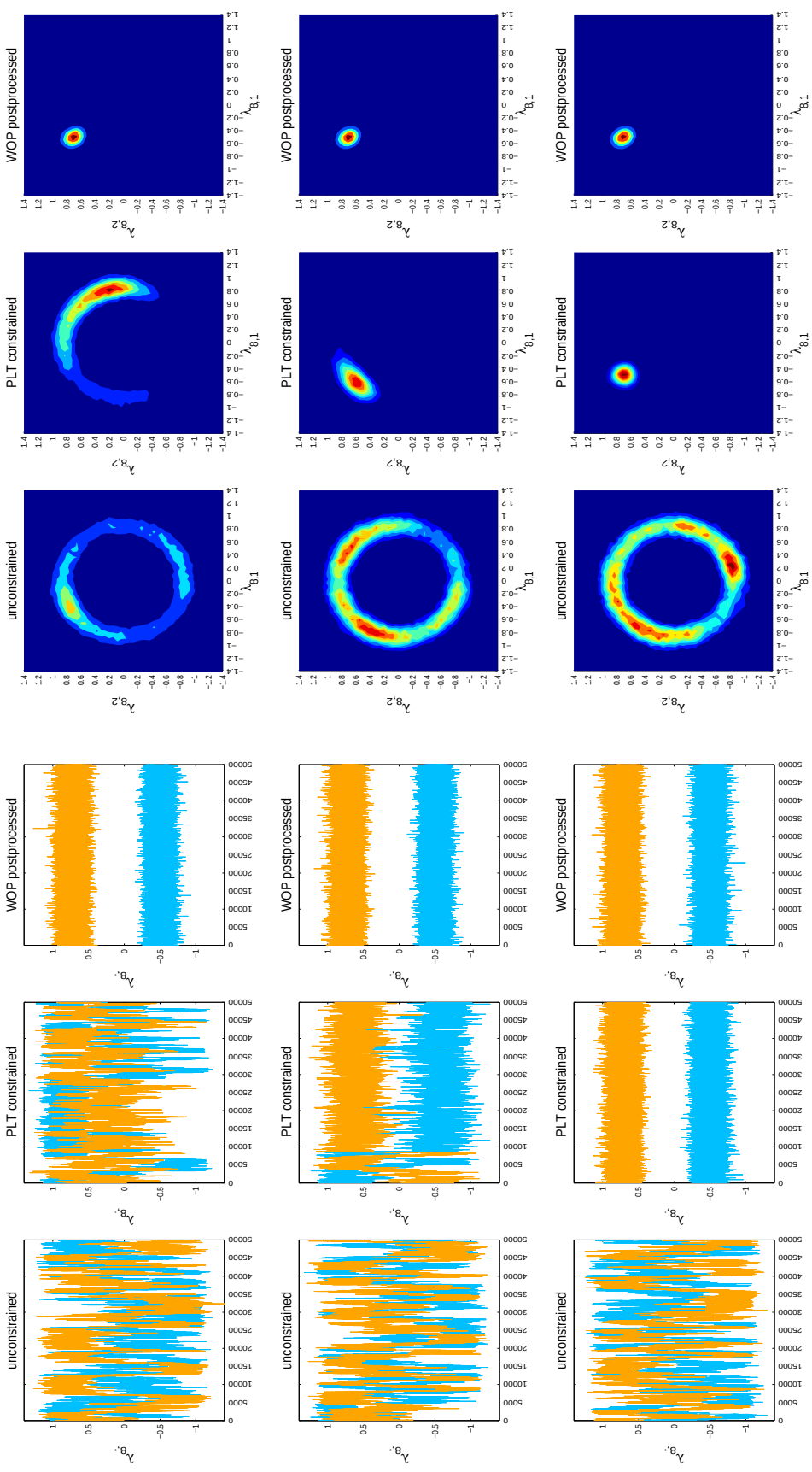

Figure 3.2: Gibbs sequences and contour plots for the bivariate posterior distributions of λ_8.

Notes: λ_8 denotes the factor loadings on variable 8, the Gibbs sequences consist of 50,000 Gibbs iterations. The plots in the top, middle and bottom row show the Gibbs sequences and posterior distributions obtained under the first, second and third ordering of the data $Y|O_1$, $Y|O_2$ and $Y|O_3$, respectively. The first and fourth column show the Gibbs sequences and posterior distributions from the unconstrained sampler, the second and fifth column show the Gibbs sequences and posterior distributions from the sampler with PLT constraints, and the third and sixth column show the Gibbs sequences and posterior distributions from the unconstrained sampler postprocessed with WOP.

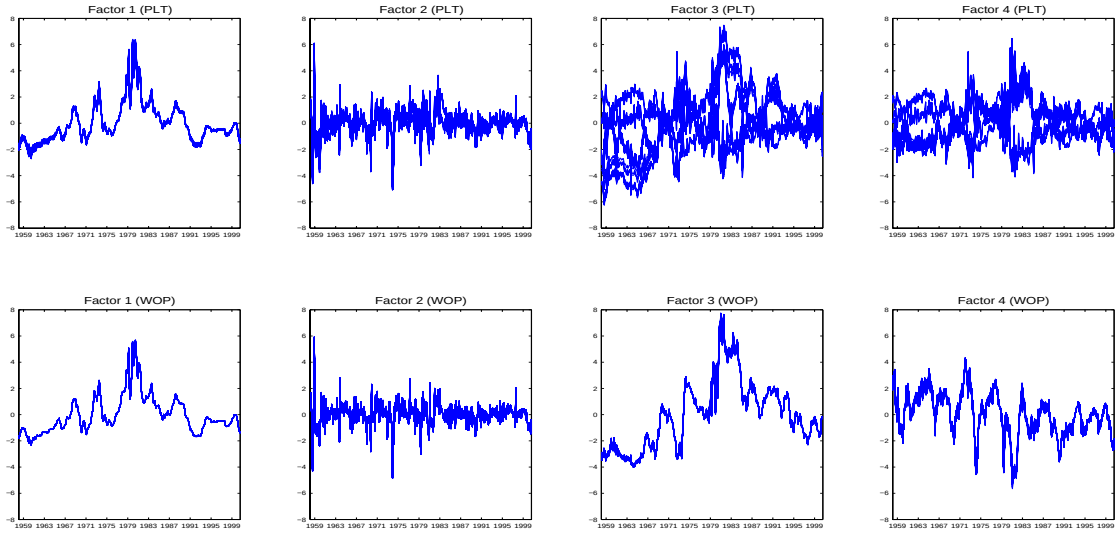

Figure 3.3: Estimated factors from 120 macroeconomic time series, displaying the results 20 randomly chosen converged sequences.

Notes: Variables chosen as factor founders are federal funds rate (FYFF), industrial production (IP), monetary base (FM2), and NAPM commodity price index (PMCP). The first row shows the results from the PLT approach, the second row shows the results from the WOP approach, which have been orthogonally transformed to obtain the same tridiagonal loadings structure as PLT.

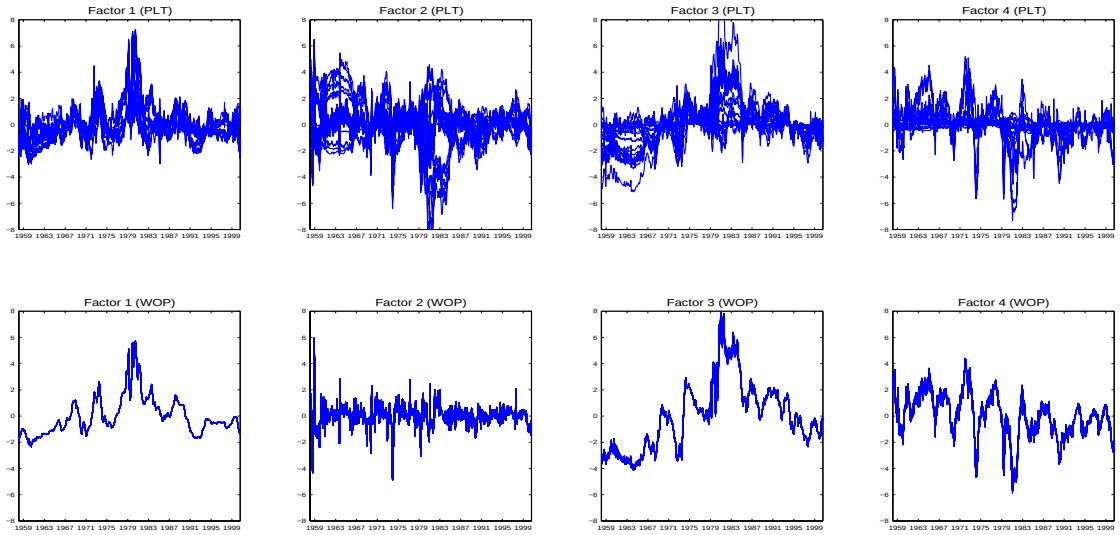

Figure 3.4: Estimated factors from 120 macroeconomic time series, displaying the results 20 randomly chosen converged sequences.

Notes: Factor founders have been set randomly, results have afterwards been orthogonally transformed to create a positive lower triangular loadings matrix on the same four variables used as factor founders before, i.e. federal funds rate (FYFF), industrial production (IP), monetary base (FM2), and NAPM commodity price index (PMCP). The first row shows the results from the PLT approach, the second row shows the results from the WOP approach.

Appendix 3.A : The Unconstrained Gibbs Sampler

For the model described in Equations (3.1) and (3.2) and prior distributions given in Equations (3.8) to (3.10), the unconstrained sampler proceeds by iteratively sampling from the corresponding full conditional distributions, see also Bai and Wang (2012).

Sampling the Latent Factors by Forward-Filtering Backward-Sampling Using a Square-Root Kalman Filter

The latent dynamic factors are obtained via forward-filtering backward-sampling, using the ensemble-transform Kalman square-root filter (ETKF) in order to improve the performance of the sampling approach, see Tippett et al. (2003). Let

$$C = \max\{P, S+1\} \tag{3.32}$$

and define

$$G = \begin{pmatrix} \Phi_1 & \dots & \Phi_{C-1} & \Phi_C \\ I_K & & 0_K & 0_K \\ & \ddots & & \vdots \\ 0_K & & I_K & 0_K \end{pmatrix} \tag{3.33}$$

as the $CK \times CK$ extended block companion matrix of the latent dynamic factors, where $\Phi_c = 0_K$ for $c > P$,

$$E_t = [\epsilon_t' \;\; 0_{1\times(C-1)K}]' \tag{3.34}$$

as the vector of error terms in the state equation,

$$Q = \begin{pmatrix} I_K & 0_{K\times(C-1)K} \\ 0_{(C-1)K\times K} & 0_{(C-1)K} \end{pmatrix} \tag{3.35}$$

as the corresponding covariance matrix, and

$$F_t = [f_t', \dots, f_{t-C}']' \tag{3.36}$$

as a vector of stacked latent factors containing the contemporary factors and C lags.[30] The state equation of the model then is obtained as

$$F_t = GF_{t-1} + E_t. \tag{3.37}$$

[30] Assume that $f_t = 0_{K\times 1}$ for $t \leq 0$ throughout.

Accordingly, the observation equation is

$$y_t = HF_t + e_t, \tag{3.38}$$

where

$$H = [\Lambda_0, \ldots, \Lambda_C], \tag{3.39}$$

with $\Lambda_c = 0_{N\times K}$ for $c > S$. With the state estimate at time $t-1$ being $\hat{F}_{t-1|t-1}$, where $\hat{F}_{0|0} = 0_{N\times 1}$, the predicted state at time t is

$$\hat{F}_{t|t-1} = G\hat{F}_{t-1|t-1}, \tag{3.40}$$

and the prediction covariance is

$$\hat{S}_{t|t-1} = G\hat{S}_{t-1|t-1}G' + Q, \tag{3.41}$$

where $\hat{S}_{0|0} = I_K$. Taking the observed value y_t into account, we obtain the prediction error

$$u_{t|t-1} = y_t - H\hat{F}_{t|t-1}. \tag{3.42}$$

The Kalman gain is obtained as

$$K_t = \hat{S}_{t|t-1}H'(H\hat{S}_{t|t-1}H + \Sigma)^{-1}, \tag{3.43}$$

hence the updating of the covariance matrix can be written as

$$\hat{S}_{t|t} = (I - K_tH)\hat{S}_{t|t-1}. \tag{3.44}$$

For the according updating step of the ETKF, we first perform a singular-value decomposition of $\hat{S}_{t|t-1}$ as

$$\hat{S}_{t|t-1} = A_{t|t-1}Z_{t|t-1}A'_{t|t-1}, \tag{3.45}$$

and define the square root of the prediction covariance as

$$Z^f_t = A_{t|t-1}Z^{\frac{1}{2}}_{t|t-1}. \tag{3.46}$$

Considering the according singular-value decomposition of the innovation covariance matrix as

$$\hat{S}_{t|t} = A_{t|t}Z_{t|t}A'_{t|t}, \tag{3.47}$$

the corresponding square root can be defined as

$$Z^a_t = A_{t|t}Z^{\frac{1}{2}}_{t|t}. \tag{3.48}$$

The result from Equation (3.46) can be inserted into Equation (3.44) to obtain

$$\hat{S}_{t|t} = Z_t^f (I - Z_t^{f'} H' (H Z_t^f Z_t^{f'} H' + \Sigma)^{-1} H Z_t^f) Z_t^{f'}, \tag{3.49}$$

hence obtaining a square root of the term in parentheses by an according singular value decomposition of an equivalent expression by the Sherman-Morrison Woodbury identity,

$$(I + Z_t^{f'} H' \Sigma^{-1} H Z_t^f)^{-1} = B_t \Gamma_t B_t', \tag{3.50}$$

or, equivalently,

$$I + Z_t^{f'} H' \Sigma^{-1} H Z_t^f = B_t \Gamma_t^{-1} B_t'. \tag{3.51}$$

The required square root is then

$$M_t = B_t \Gamma_t^{-\frac{1}{2}}, \tag{3.52}$$

allowing for the square-root updating as

$$Z_t^a = Z_t^f M_t. \tag{3.53}$$

Then the innovation covariance matrix can be rebuilt as

$$\hat{S}_{t|t} = Z_t^a Z_t^{a\prime}, \tag{3.54}$$

and the updated mean is

$$\hat{F}_{t|t} = \hat{F}_{t|t-1} + \hat{S}_{t|t} H \Sigma^{-1} u_{t|t-1}. \tag{3.55}$$

The factors are then obtained by backward-sampling from the resulting $\hat{F}_{t|T}$ and $\hat{S}_{t|T}$.

The Remaining Parameters

Throughout the paper, we assume diagonality for Σ resulting in

$$f(\Sigma | Y, \{\Lambda_s\}_{s=0}^S, \{\Phi_p\}_{p=1}^P, \{f_t\}_{t=1}^T) = \prod_{i=1}^N \frac{b_i^{a_i}}{\Gamma(a_i)} \left(\frac{1}{\sigma_i^2}\right)^{a_i - 1} \exp\left\{-\frac{1}{\sigma_i^2} b_i\right\}, \tag{3.56}$$

where $a_i = \frac{1}{2}T + \underline{\alpha}_i$ and $b_i = \frac{1}{2}\sum_{t=1}^T (y_{it} - \sum_{s=0}^S \lambda_{s,i}' f_{t-s})^2 + \underline{\beta}_i$ and $\underline{\alpha}_i = \underline{\beta}_i = 1$ for all $i = 1 \ldots, N$. Due to diagonality of Σ, the full conditional distribution of the loadings can be factorized over the $S + 1$ Λ_s matrices, and row-wise within these matrices, taking the

N individual rows $\lambda_{s,i}$ per matrix into account. This yields the following full conditional distribution:

$$f(\{\Lambda_s\}_{s=0}^{S}|Y,\Sigma,\{\Phi_p\}_{p=1}^{P},\{f_t\}_{t=1}^{T}) = \prod_{s=0}^{S}\prod_{i=1}^{N}(2\pi)^{-\frac{K}{2}}|\Omega_{\lambda_{s,i}}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\lambda_{s,i}-\mu_{\lambda_{s,i}})'\Omega_{\lambda_{s,i}}^{-1}(\lambda_{s,i}-\mu_{\lambda_{s,i}})\right\}, \tag{3.57}$$

where $\Omega_{\lambda_{s,i}} = (\frac{1}{\sigma_i^2}\sum_{t=1}^{T} f_{t-s}f_{t-s}' + (\underline{\Upsilon}_s)_{i,i}\, I_K)^{-1}$ and $\mu_{\lambda_{s,i}} = \Omega_{\lambda_{s,i}}(\frac{1}{\sigma_i^2}\sum_{t=1}^{T} y_{it}f_{t-s}')$.

Finally, consider a stacked version of the persistence parameters for the factors,

$$\tilde{\Phi} = [\Phi_1',\ \ldots,\ \Phi_P']' \tag{3.58}$$

and denote a shortened $T-P\times K$ factor matrix starting at time point t as

$$\tilde{F}_t = [f_t,\ \ldots,\ f_{T-P+(t-1)}]', \tag{3.59}$$

and

$$\tilde{F} = [\tilde{F}_1,\ \ldots,\ \tilde{F}_P] \tag{3.60}$$

containing P such matrices. Then the full conditional distribution of $\tilde{\Phi}$ for normally distributed innovations in the factors and with an uninformative prior distribution obtains as

$$f(\tilde{\Phi}|Y,\Sigma,\{\Lambda_s\}_{s=0}^{S},\{f_t\}_{t=1}^{T}) = (2\pi)^{-\frac{KP}{2}}|\Omega_{\tilde{\Phi}}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathrm{vec}(\tilde{\Phi})-\mu_{\tilde{\Phi}})'\Omega_{\tilde{\Phi}}{}^{-1}(\mathrm{vec}(\tilde{\Phi})-\mu_{\tilde{\Phi}})\right\}, \tag{3.61}$$

where $\Omega_{\tilde{\Phi}} = \Psi_\epsilon^{-1}\otimes(\tilde{F}'\tilde{F})^{-1}$ and $\mu_{\tilde{\Phi}} = \mathrm{vec}((\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{F}_{P+1})$, see e.g. Ni and Sun (2005).

Appendix 3.B: Proof of Proposition 3.3.2

Proof. Given a parametrization of D ensuring orthogonality the minimization problem in Equation (3.22) can be restated as

$$D = \arg\max \mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^{*}) + \mathrm{tr}\left(\sum_{p=1}^{P} D'\Phi_p^{(z)\prime}D\Phi_p^{*}\right). \tag{3.62}$$

To start with, let $K=2$ and look at $\mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^{*})$ first. Define

$$M = \bar{\Lambda}^{(z)\prime}\bar{\Lambda}^{*}, \tag{3.63}$$

and assume the parametrization $D = D_+$, i.e.

$$D_+ = \begin{pmatrix} \cos(\gamma_+) & -\sin(\gamma_+) \\ \sin(\gamma_+) & \cos(\gamma_+) \end{pmatrix}. \tag{3.64}$$

Then D_+ can be expressed in terms of an angle $\gamma_+ \in [-\pi, \pi)$ resulting in

$$\begin{aligned} \operatorname{tr}(D_+' M) &= \operatorname{tr} \begin{pmatrix} m_{11}\cos(\gamma_+) + m_{21}\sin(\gamma_+) & m_{12}\cos(\gamma_+) + m_{22}\sin(\gamma_+) \\ -m_{11}\sin(\gamma_+) + m_{21}\cos(\gamma_+) & -m_{12}\sin(\gamma_+) + m_{22}\cos(\gamma_+) \end{pmatrix} \\ &= (m_{11} + m_{22})\cos(\gamma_+) + (m_{21} - m_{12})\sin(\gamma_+) \\ &= \sqrt{(m_{11} + m_{22})^2 + (m_{21} - m_{12})^2}\cos(\gamma_+ - \operatorname{atan2}(m_{11} + m_{22}, m_{21} - m_{12})) \\ &= A_+ \cos(\gamma_+ + \varphi_+), \end{aligned} \tag{3.65}$$

which is a sinusoid, see e.g. Shumway and Stoffer (2010), Chapter 4.2, with amplitude

$$A_+ = \sqrt{(m_{11} + m_{22})^2 + (m_{21} - m_{12})^2}, \tag{3.66}$$

phase $\varphi_+ = -\operatorname{atan2}(m_{11} + m_{22}, m_{21} - m_{12})$, and frequency $\omega = \frac{1}{2\pi}$, i.e. there is exactly one maximum in the domain of γ for any choice of D_+.[31] Note that (3.65) uses the important equality

$$A\cos(\omega t + \varphi) = \sum_{i=1}^{n} A_i \cos(\omega t + \varphi_i), \tag{3.67}$$

where

$$A = \sqrt{\left(\sum_{i=1}^{n} A_i \cos(\varphi_i)\right)^2 + \left(\sum_{i=1}^{n} A_i \sin(\varphi_i)\right)^2} \tag{3.68}$$

and

$$\varphi = \operatorname{atan2}\left(\sum_{i=1}^{n} A_i \cos(\varphi_i), \sum_{i=1}^{n} A_i \sin(\varphi_i)\right), \tag{3.69}$$

for which a proof can be found e.g. in Smith (2007). The resulting matrix for $K = 2$ is hence

$$D_- = \begin{pmatrix} \cos(\gamma_-) & \sin(\gamma_-) \\ \sin(\gamma_-) & -\cos(\gamma_-) \end{pmatrix}. \tag{3.70}$$

[31] The two-argument arctangent function $\operatorname{atan2}(y, x)$, defined on the interval $[-\pi, \pi)$ and based on a half-angle identity for the tangent, is given as

$$\operatorname{atan2}(y, x) = 2\arctan\frac{\sqrt{x^2 + y^2} - x}{y}.$$

Then D_- can again be expressed in terms of an angle $\gamma_- \in [-\pi, \pi)$ resulting in

$$\begin{aligned}\operatorname{tr}(D_-'M) &= \operatorname{tr}\begin{pmatrix} m_{11}\cos(\gamma_-) + m_{21}\sin(\gamma_-) & m_{12}\cos(\gamma_-) + m_{22}\sin(\gamma_-) \\ m_{11}\sin(\gamma_-) - m_{21}\cos(\gamma_-) & m_{12}\sin(\gamma_-) - m_{22}\cos(\gamma_-) \end{pmatrix} \qquad (3.71)\\ &= (m_{11} - m_{22})\cos(\gamma_-) + (m_{12} + m_{21})\sin(\gamma_-) \\ &= \sqrt{(m_{11} - m_{22})^2 + (m_{12} + m_{21})^2}\cos(\gamma_- - \operatorname{atan2}(m_{11} - m_{22}, m_{12} + m_{21})) \\ &= A_- \cos(\gamma_- + \varphi_-), \qquad (3.72)\end{aligned}$$

which is also a sinusoid, but with amplitude $A_- = \sqrt{(m_{11} - m_{22})^2 + (m_{12} + m_{21})^2}$, phase $\varphi_- = -\operatorname{atan2}(m_{11} - m_{22}, m_{12} + m_{21})$, and frequency $\omega = \frac{1}{2\pi}$. Thus, γ_- and γ_+ are uniquely identified if A_- and φ_- or A_+ and φ_+ respectively are all distinct from zero. Note that the events $m_{11} - m_{22} = 0$ and $m_{21} + m_{12} = 0$ or $m_{11} + m_{22} = 0$ and $m_{21} - m_{12} = 0$ corresponding to A_- and φ_- being zero or A_+ and φ_+ being zero respectively occur with probability zero since the corresponding restrictions on $\vartheta^{(z)}$ and ϑ^* denote a subspace of the parameter space. Further, the two maxima implied by γ_- and γ_+ are distinct with probability one since the event $A_- = A_+$ occurs as well with probability zero.

Now look at $\operatorname{tr}\left(\sum_{p=1}^{P} D'\Phi_p^{(z)\prime} D\Phi_p^*\right)$ and let $P = 1$. Assuming $D = D_+$ yields

$$\begin{aligned}\operatorname{tr}(D'\Phi^{(z)\prime}D\Phi^*) = {}& \phi_{12}^*(k_1(\gamma_+) + k_2(\gamma_+)) + \phi_{11}^*(k_3(\gamma_+) + k_4(\gamma_+)) \\ &+ \phi_{22}^*(k_5(\gamma_+) + k_6(\gamma_+)) + \phi_{21}^*(k_7(\gamma_+) + k_8(\gamma_+)), \qquad (3.73)\end{aligned}$$

with

$$\begin{aligned}
k_1(\gamma_+) &= -\cos(\gamma_+)(\phi_{12}^{(z)}\cos(\gamma_+) + \phi_{11}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{12}^{(z)}\cos^2(\gamma_+) + \phi_{11}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.74)\\
k_2(\gamma_+) &= -\sin(\gamma_+)(\phi_{22}^{(z)}\cos(\gamma_+) + \phi_{21}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{21}^{(z)}\sin^2(\gamma_+) - \phi_{22}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.75)\\
k_3(\gamma_+) &= -\cos(\gamma_+)(\phi_{11}^{(z)}\cos(\gamma_+) - \phi_{12}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{11}^{(z)}\cos^2(\gamma_+) - \phi_{12}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.76)\\
k_4(\gamma_+) &= -\sin(\gamma_+)(\phi_{21}^{(z)}\cos(\gamma_+) - \phi_{22}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{22}^{(z)}\sin^2(\gamma_+) - \phi_{21}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.77)\\
k_5(\gamma_+) &= -\cos(\gamma_+)(\phi_{22}^{(z)}\cos(\gamma_+) + \phi_{21}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{22}^{(z)}\cos^2(\gamma_+) + \phi_{21}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.78)\\
k_6(\gamma_+) &= -\sin(\gamma_+)(\phi_{12}^{(z)}\cos(\gamma_+) + \phi_{11}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{11}^{(z)}\sin^2(\gamma_+) + \phi_{12}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.79)\\
k_7(\gamma_+) &= -\cos(\gamma_+)(\phi_{21}^{(z)}\cos(\gamma_+) - \phi_{22}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{21}^{(z)}\cos^2(\gamma_+) - \phi_{22}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.80)\\
k_8(\gamma_+) &= -\sin(\gamma_+)(\phi_{11}^{(z)}\cos(\gamma_+) - \phi_{12}^{(z)}\sin(\gamma_+)) \\
&= -\phi_{12}^{(z)}\sin^2(\gamma_+) + \phi_{11}^{(z)}\sin(\gamma_+)\cos(\gamma_+) && (3.81)
\end{aligned}$$

Consider Equations (3.75), (3.77), (3.79) and (3.81) and obtain

$$\begin{aligned}
k_2 &= -\phi_{21}^{(z)}\sin^2(\gamma_+) - \phi_{22}^{(z)}\sin(\gamma_+)\cos(\gamma_+) \\
&= -\phi_{21}^{(z)}(1-\cos^2(\gamma_+)) - \phi_{22}^{(z)}\sin(\gamma_+)\cos(\gamma_+) = k_7 - \phi_{21}^{(z)} && (3.82)\\
k_4 &= -\phi_{22}^{(z)}\sin^2(\gamma_+) - \phi_{21}^{(z)}\sin(\gamma_+)\cos(\gamma_+) \\
&= -\phi_{22}^{(z)}(1-\cos^2(\gamma_+)) - \phi_{21}^{(z)}\sin(\gamma_+)\cos(\gamma_+) = \phi_{22}^{(z)} - k_5 && (3.83)\\
k_6 &= -\phi_{11}^{(z)}\sin^2(\gamma_+) + \phi_{12}^{(z)}\sin(\gamma_+)\cos(\gamma_+) \\
&= -\phi_{11}^{(z)}(1-\cos^2(\gamma_+)) + \phi_{12}^{(z)}\sin(\gamma_+)\cos(\gamma_+) = \phi_{11}^{(z)} - k_3 && (3.84)\\
k_8 &= -\phi_{12}^{(z)}\sin^2(\gamma_+) + \phi_{11}^{(z)}\sin(\gamma_+)\cos(\gamma_+) \\
&= -\phi_{12}^{(z)}(1-\cos^2(\gamma_+)) + \phi_{11}^{(z)}\sin(\gamma_+)\cos(\gamma_+) = k_1 - \phi_{12}^{(z)} && (3.85)
\end{aligned}$$

Inserting (3.74), (3.76), (3.78), (3.80), (3.82), (3.83), (3.84) and (3.85) into (3.73) yields

$$\begin{aligned}
\operatorname{tr}(D'\Phi^{(z)\prime}D\Phi^*) &= \phi_{12}^*(k_1+k_7-\phi_{21}^{(z)}) + \phi_{11}^*(k_3-k_5+\phi_{22}^{(z)}) \\
&\quad + \phi_{22}^*(k_5-k_3+\phi_{11}^{(z)}) + \phi_{21}^*(k_7+k_1-\phi_{12}^{(z)}) \\
&= \underbrace{-\phi_{12}^*\phi_{21}^{(z)} - \phi_{11}^*\phi_{22}^{(z)} + \phi_{22}^*\phi_{11}^{(z)} - \phi_{21}^*\phi_{12}^{(z)}}_{=c_0} + \\
&\quad (\phi_{12}^*+\phi_{21}^*)(k_1+k_7) + (\phi_{11}^*-\phi_{22}^*)(k_3-k_5) \\
&= c_0 + (\phi_{12}^*+\phi_{21}^*)((\phi_{12}^{(z)}+\phi_{21}^{(z)})\cos^2(\gamma_+) + (\phi_{11}^{(z)}-\phi_{22}^{(z)})\sin(\gamma_+)\cos(\gamma_+)) \\
&\quad + (\phi_{11}^*-\phi_{22}^*)((\phi_{11}^{(z)}-\phi_{22}^{(z)})\cos^2(\gamma_+) - (\phi_{12}^{(z)}+\phi_{21}^{(z)})\sin(\gamma_+)\cos(\gamma_+)) \\
&= c_0 + \underbrace{((\phi_{12}^*+\phi_{21}^*)(\phi_{12}^{(z)}+\phi_{21}^{(z)}) + (\phi_{11}^*-\phi_{22}^*)(\phi_{11}^{(z)}-\phi_{22}^{(z)}))}_{=c_1}\cos^2(\gamma_+) \\
&\quad + \underbrace{((\phi_{12}^*+\phi_{21}^*)(\phi_{11}^{(z)}-\phi_{22}^{(z)}) - (\phi_{11}^*-\phi_{22}^*)(\phi_{12}^{(z)}+\phi_{21}^{(z)}))}_{=c_2}\sin(\gamma_+)\cos(\gamma_+) \\
&= c_0 + (c_1\cos(\gamma_+) + c_2\sin(\gamma_+))\cos(\gamma_+) \\
&= c_0 + (\underbrace{\sqrt{c_1^2+c_2^2}}_{=c_3}\cos(\gamma_+ + \underbrace{\operatorname{atan2}(c_1,c_2)}_{=c_4}))\cos(\gamma_+) \\
&= c_0 + \frac{1}{2}c_3\cos(c_4) + \frac{1}{2}c_3\cos(2\gamma_+ - c_4) \\
&= V_+ + A_+\cos(2\gamma_+ + \varphi_+), && (3.86)
\end{aligned}$$

where c_0 through c_4 are constant terms, and the second-last equality uses the fact that

$$\cos(\gamma_1)\cos(\gamma_2) = \frac{1}{2}\left(\cos(\gamma_1-\gamma_2) + \cos(\gamma_1+\gamma_2)\right). \tag{3.87}$$

The result of (3.86) is sinusoid with vertical shift $V_+ = c_0 + \frac{1}{2}c_3\cos(c_4)$, amplitude $A_+ = \frac{1}{2}c_3$, phase $\varphi_+ = -c_4$, and frequency $\omega = \frac{1}{\pi}$, i.e. there are exactly two maxima in the domain of γ for any choice of D_+. The equivalent result for $D = D_-$ obtains analogously. For $P > 1$, reversing the order of summation and trace operator in $\operatorname{tr}\left(\sum_{p=1}^{P} D'\Phi_p^{(z)\prime}D\Phi_p^*\right)$, we obtain P such sinusoids, which all depend on the same γ_+, thus we can apply Equation (3.67) to

the demeaned sinusoids and afterwards add the sum of the means again, another vertically shifted sinusoid with frequency $\frac{1}{\pi}$ and thus two maxima in the domain of γ_+. Note that the choice of $D = D_+$ yields the superposition of P sinusoids, whereas the choice of $D = D_-$ yields another superposition of P sinusoids, however with according changes in the phase, amplitude and vertical shift parameters. Although for D_+ as well as for D_- we find two maxima each, the maxima under the two parametrization are distinct with probability one as the restrictions on the parameter space causing coincidence of the two sets of maxima under the two parametrizations of the orthogonal matrix refer to a subspace of the parameter space having thus probability zero. Further using the same line of argument as above, for each of the parametrizations there exist two maxima with probability one.

To show the uniqueness of the maximum of $\mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*) + \mathrm{tr}\left(\sum_{p=1}^{P} D'\Phi_p^{(z)\prime} D\Phi_p^*\right)$ we must consequently consider for both cases, $D = D_+$ and $D = D_-$, a superposition of two sinusoids with frequency $\frac{1}{2\pi}$ and $\frac{1}{\pi}$, respectively. The first of them has one peak and one trough on the interval $[-\pi, \pi)$, while the second has two of each. The sum over these two sinusoids has two peaks of identical height if and only if the peak of the first coincides with one of the two troughs of the second. Denoting the phase of $\mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)$ as φ_Λ and the phase of $\mathrm{tr}\left(\sum_{p=1}^{P} D'\Phi_p^{(z)\prime} D\Phi_p^*\right)$ as φ_Φ this implies the strict equality $\varphi_\Phi = \pi + 2\varphi_\Lambda$ corresponding to a restriction of the parameter space having probability zero.

Now consider the general case for $K > 2$. To derive the structure of the expression $\mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)$, look at the matrix D first. D can be expressed as the product over $K(K-1)/2$ Givens rotation matrices and a reflection about the K^{th} axis. For the time being, the reflection is not considered. The Givens rotation matrices are functions in the angles $\underline{\gamma} = (\gamma_1, \ldots, \gamma_{\frac{K(K-1)}{2}})$. Thus, defining the constituent set of elements

$$\mathcal{CS} = \left\{\cos(\gamma_{k^*}), \sin(\gamma_{k^*}) = \cos\left(\gamma_{k^*} - \frac{\pi}{2}\right)\right\}_{k^*=1}^{K(K-1)/2}, \tag{3.88}$$

each entry of D can be characterized as

$$d_{ij} = \sum_{j^*=1}^{T_{ij}} a_{j^*}^{ij} \prod_{k^*=1}^{K(K-1)/2} \cos(\gamma_{k^*})^{b_{j^*k^*}^{ij}} \cos\left(\gamma_{k^*} - \frac{\pi}{2}\right)^{c_{j^*k^*}^{ij}}, \tag{3.89}$$

with T_{ij} denoting the number of subsets involved in d_{ij}, $a_{j^*}^{ij} \in \{-1, 1\}$, $b_{j^*k^*}^{ij}$ and $c_{j^*k^*}^{ij}$ taking either values 0 or 1, and $b_{j^*k^*}^{ij} + c_{j^*k^*}^{ij} \leq 1$. Then

$$\mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*) = \sum_{j=1}^{K} d'_{\cdot j}(\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)_{\cdot j}, \tag{3.90}$$

where $D_{\cdot j}$ denotes the j^{th} column of D and $(\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)_{\cdot j}$ denotes the j^{th} column of $\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*$. The same expression can also be stated in the structural form from Equation (3.89), hence

$$\operatorname{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*) = \sum_{j^*=1}^{T_{\text{tr}\bar{\Lambda}}} q_{j^*}^{\text{tr}\bar{\Lambda}} \prod_{k^*=1}^{K(K-1)/2} \cos(\gamma_{k^*})^{b_{j^*k^*}^{\text{tr}\bar{\Lambda}}} \cos\left(\gamma_{k^*} - \frac{\pi}{2}\right)^{c_{j^*k^*}^{\text{tr}\bar{\Lambda}}}, \tag{3.91}$$

where $T_{\text{tr}\bar{\Lambda}}$ denotes the number of subsets entering $\operatorname{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)$, $q_{j^*}^{\text{tr}\bar{\Lambda}}$ is a function of the $a_{j^*}^{ij}$ and the elements in the matrix $\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*$, and $b_{j^*k^*}^{\text{tr}\bar{\Lambda}}$ and $c_{j^*k^*}^{\text{tr}\bar{\Lambda}}$ taking either values 0 or 1, and $b_{j^*k^*}^{\text{tr}\bar{\Lambda}} + c_{j^*k^*}^{\text{tr}\bar{\Lambda}} \leq 1$. It can be seen from Equation (3.90) that each subset involved in any d_{ij} enters Equation (3.91), which is hence a weighted sum over the union of products of all subsets of $\mathcal{CS}$ involved in D.

A weighted one-element subset of $\mathcal{CS}$ is a sinusoid with frequency $\frac{1}{2\pi}$, as discussed for $K = 2$. Since the γ_{k^*} are all mutually independent, the multiple-element subsets of $\mathcal{CS}$ are therefore sinusoids with the same frequency along each dimension and dimensionality not larger than $K(K-1)/2$. $\operatorname{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)$ is then the superposition of $T_{\text{tr}\bar{\Lambda}}$ such sinusoids. In fact, all sinusoids can be treated as $K(K-1)/2$-variate sinusoids, which are constant along the dimensions whose angles they do not depend on. The superposition yields a unique maximum if each γ_{k^*} enters at least one of the sinusoids, otherwise the value of the respective γ_{k^*} is irrelevant for the maximization. Further, it must be ensured that all sinusoids have a unique maximum in the subset of $\underline{\gamma}$ they depend on, or, if this is not the case, the different parametrizations in $\underline{\gamma}$ all imply the same D.

Unlike the univariate sinusoids with frequency $\frac{1}{2\pi}$, however, multivariate ones have multiple maxima, because joint replacements of pairs of elements of $\underline{\gamma}$ can exploit the trigonometric identity

$$\cos(\gamma) = -\cos(\gamma + \pi) = -\cos(\gamma - \pi) = \cos(-\gamma). \tag{3.92}$$

Without loss of generality, consider the bivariate sinusoid $\cos(\gamma_1)\cos(\gamma_2)$, which has a maximum in $\underline{\gamma} = (0, 0)$. By Equation (3.92), the exists a second maximum in $\underline{\gamma} = (\pi, \pi)$. The case of the bivariate sinusoid is also shown in the left panel of Figure 3.5. Accordingly, a trivariate sinusoid allows for $\binom{3}{2} = 3$ pairwise replacements, and a 4-variate sinusoid allows for $\binom{4}{2} + \binom{4}{4} = 6 + 1 = 7$ replacements, where the second term denotes the replacement of two pairs of angles by their counterparts at the same time. The number of additional redundant parametrizations for a $K(K-1)/2$-variate sinusoid is thus $\sum_{i^*=1}^{\lfloor \frac{K(K-1)}{2}/2 \rfloor} \binom{\frac{K(K-1)}{2}}{2i^*}$, implying a total number of modes of $2^{\frac{K(K-1)}{2}-1}$. Note, however, that in order to obtain a redundant parametrization of D, *all* involved sinusoids must allow for the according pairwise replacements. The actual number of modes is therefore usually much smaller than $2^{\frac{K(K-1)}{2}-1}$. Consider e.g. $K = 3$, where the only admissible replacement for $\underline{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ is $\tilde{\underline{\gamma}} = (\gamma \pm \pi, \pm\pi - \gamma_2, \gamma_3 \pm \pi)$, where the sign of $\pm\pi$ must be chosen such that the angle is in

the admissible range for $\underline{\gamma}$. Taking the redundant parametrizations into account, there exists thus a unique orthogonal matrix D providing a maximum for the involved sinusoids and thus a unique D maximizing $\mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)$.

An expression analogous to the one in Equation (3.91) can also be found for $\sum_{p=1}^{P}\mathrm{tr}(D'\Phi_p^{(z)\prime}D\Phi_p^*)$. Note that here, it is possible that the $\cos(\gamma_{k^*})$ enter in quadratic form, hence, the resultant sinusoids have frequency $\frac{1}{\pi}$. $\sum_{p=1}^{P}\mathrm{tr}(D'\Phi_p^{(z)\prime}D\Phi_p^*)$ then has the structural form

$$\sum_{j^*=1}^{T_{\mathrm{tr}\Phi}} p_{j^*}^{\mathrm{tr}\Phi} \prod_{k^*=1}^{K(K-1)} \cos(\gamma_{k^*})^{b_{j^*k^*}^{\mathrm{tr}\Phi}} \cos\left(\gamma_{k^*}-\frac{\pi}{2}\right)^{c_{j^*k^*}^{\mathrm{tr}\Phi}}, \tag{3.93}$$

with $b_{j^*k^*}^{\mathrm{tr}\Phi}$ and $c_{j^*k^*}^{\mathrm{tr}\Phi}$ taking values $\{0,1,2\}$, $b_{j^*k^*}^{\mathrm{tr}\Phi}+c_{j^*k^*}^{\mathrm{tr}\Phi}\leq 2$, where $p_{j^*}^{\mathrm{tr}\Phi}$ is a function of the elements involved in the matrices $\Phi_p^{(z)}$ and Φ_p^*, $p=1,\ldots,P$. Hence $\sum_{p=1}^{P}\mathrm{tr}(D'\Phi_p^{(z)\prime}D\Phi_p^*)$ is the sum of sinusoids having frequency $\frac{1}{\pi}$ or $\frac{1}{2\pi}$ along each dimension. Consequently, assuming that all γ_{k^*} enter the expression in Equation (3.93) at least once, the result is a superposition of $K(K-1)/2$-variate sinusoids, which do not exceed the frequency $\frac{1}{\pi}$ in any dimension. A bivariate sinusoid with frequency $\frac{1}{\pi}$ in each dimension is shown in the right panel of Figure 3.5. Each dimension where the frequency is doubled necessarily has twice as many maxima. Nonetheless, the number of maxima cannot exceed $2\cdot 4^{\frac{K(K-1)}{2}-1}$ and is thus finite. Superimposing the sinusoids in $\sum_{j^*=1}^{T_{\mathrm{tr}\Phi}} p_{j^*}^{ij}$ with those in $\mathrm{tr}(D'\bar{\Lambda}^{(z)\prime}\bar{\Lambda}^*)$ thus results in a unique maximum almost surely, where the event that two maxima of the superimposed sinusoids are equally qualified by the sinusoids with lower frequency corresponds to a restriction on the parameter space and hence occurs with probability zero. The same maximization over $\underline{\gamma}$, but involving a reflection over the K^{th} axis, yields a lower or higher value with probability one. In the latter case, the corresponding matrix D with $\det(D)=-1$ yields the unique maximum, in the former case, the matrix D with $\det(D)=1$, not involving the axis reflection, yields the unique maximum. □

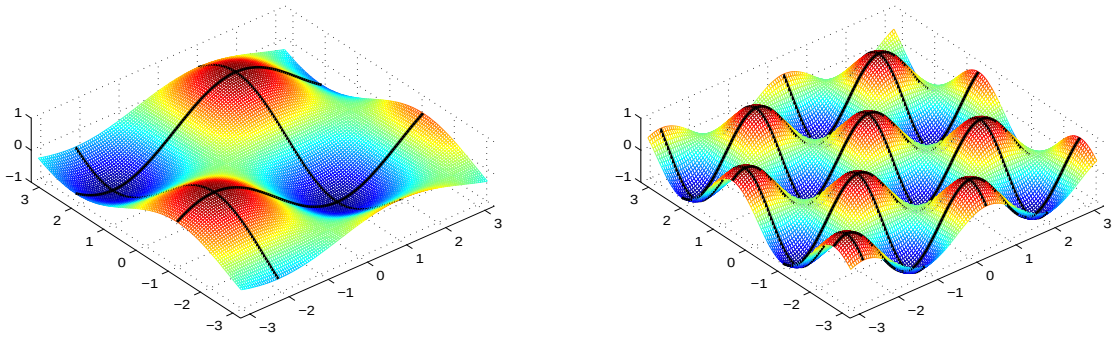

Figure 3.5: Bivariate sinusoids with frequency $\frac{1}{2\pi}$ along each dimension (left) and frequency $\frac{1}{\pi}$ along each dimension (right).

Chapter 4

A Two-Step Approach to Bayesian Analysis of Sparse Factor Models

An earlier version of this chapter was made available as an SSRN Working Paper on February 14, 2014, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2399368.

4.1 Introduction

Sparse factor analysis, see e.g. West (2003), comprises aspects of exploratory and confirmatory factor analysis, seeking to establish a parsimonious structure in the loadings matrix of the model. This task is related to the issue of determining the number of factors required for model representation, see e.g. Bai and Ng (2002), Hallin and Liska (2007) or Breitung and Pigorsch (2013), the question of which variables are relevant for the analysis and which can be excluded, see e.g. Boivin and Ng (2006), and, of course, the problem whether some variables are driven by a subset of all factors only. Whereas sparsity analysis focuses mainly on the third of these questions, it can provide helpful hints to tackle the first two questions as well. Sparse modeling is helpful particularly if very large amounts of data are at hand, such as in gene expression analysis, which has seen a host of applications and model refinements, see e.g. Carvalho et al. (2008); Lucas et al. (2006). In the field of economics, there have recently been several applications for large macroeconomic data sets, see e.g. Francis et al. (2012), Kaufmann and Schumacher (2012) and Kaufmann and Schumacher (2013).

The Bayesian inference approaches to sparse factor analysis by West (2003) and Lucas et al. (2006) use MCMC techniques, in which a sample from the posterior distribution of the model parameters and factors is generated. The sampler that is used resembles the Gibbs samplers for static and dynamic factor models described in Chapter 2 and 3, however, it uses a spike and slab prior for the factor loadings, which is a mixture of a Dirac delta distribution and a normal distribution.[1] The use of this prior results in the posterior distribution of the factor loadings likewise being a mixture, where the probabilities for the loadings being zero, or the *association probabilities* between a particular factor and a particular variable, are generally

[1] The use of this prior requires two Metropolis steps per iteration in the sampling approach, which is hence not a Gibbs sampler in the strict sense anymore.

either close to zero or close to one. The spike and slab prior is originally implemented as a single-layer prior in West (2003), implying that the association probabilities per factor are the same for all variables, and extended to a two-layer prior by Lucas et al. (2006), allowing for variable-specific association probabilities per factor. The latter approach, discussed in detail in Carvalho (2006), and slightly altered by Kaufmann and Schumacher (2013) is used as a reference in this chapter.

Due to the fact that the sampler sets some of the loadings per factor to zero, the sampled loadings matrices look similar to those from confirmatory factor analysis. In fact, confirmatory factor analysis can be seen as a special case of sparse factor analysis with association probabilities fixed either to zero or to one throughout. Some findings from confirmatory factor analysis may therefore apply to sparse factor analysis as well. In confirmatory factor analysis, statistical and numerical issues have been observed that affect inference in various ways. One possible statistical issue is the existence of multiple solutions that are all in line with the distributional assumptions of the model and the postulated association structure between the factors and variables. Dunn (1973) and Jennrich (1978) discuss cases where multiple solutions for a particular structure of the loadings matrix exist, and Bekker (1986) gives necessary and sufficient conditions for the structure such that a unique solution exists. Moreover, numerical issues have been reported in confirmatory factor analysis e.g. by Millsap (2001), Loken (2005) and Erosheva and Curtis (2013). These issues result from the shape of the likelihood under the postulated association structure, which may make it hard to find the global mode in maximum likelihood factor analysis and which may obstruct the mixing of the sampler in Bayesian factor analysis. The sampler may thus display spurious convergence, indicating that only a single mode of the posterior distribution exists, leaving additional modes undiscovered.

The attempt to find a parsimonious loadings structure in sparse factor models is closely related to the attempt to find a *simple structure* or *simple configuration* of the loadings matrix, as has been suggested by Thurstone (1938). This is typically achieved by rotating the parameter estimates conditional on specific criteria, of which the most popular is the Varimax criterion by Kaiser (1958). The novel procedure I propose follows this idea, but uses the outcome of a Bayesian inference approach, a Gibbs sampler without rotational identification constraints, where the Gibbs output is postprocessed with the weighted orthogonal Procrustes (WOP) ex-post identification approach as described in Chapter 3. This postprocessed Gibbs output contains information about the posterior density of the parameters beyond the point estimates of the parameters that the Varimax criterion is usually applied to. Hence it allows to construct highest posterior density intervals (HPDIs), or highest posterior density ellipsoids (HPDEs), for the loadings parameters.[2] Construction of the HPDEs becomes especially easy due to the fact that the marginal posterior densities of each row vector of the loadings matrix are elliptically shaped and can be achieved by means of an approach by Hanson and McMillan (2012), or by directly calculating the contours of the ellipsoids. As the WOP-postprocessed sampler output is invariant to joint orthogonal transformations, it is possible to find orthogonal

[2] An ellipsoid is understood here as the K-dimensional generalization of an ellipse, where $K \geq 2$.

transformations of the HPDEs implying either a maximally parsimonious structure in the loadings matrix or a parsimonious structure that is easy to interpret. Unlike the standard approaches to sparse factor analysis, there is no danger of the sampler "locking in" to a particular sparse structure in the loadings matrix while leaving other possible sparse structures unconsidered.

I test the proposed procedure in a simulation study, which uses simulated data according to a specification given in Lopes and West (2004) and Frühwirth-Schnatter and Lopes (2012), attempting to answer the aforementioned three questions about the number of factors, the relevant subset of variables and an appropriate sparse loadings structure, and at the comparing its performance to that of the MCMC approach by Carvalho (2006). Afterwards, I apply the novel procedure to the well-known students test data set by Holzinger and Swineford (1939), which has been used to illustrate the bi-factor model proposed in Holzinger and Swineford (1937) and is also a frequently used example in the identification of a simple structure. For reference, the same data is also analyzed by classical principal components (PC) factor analysis with a subsequent Varimax rotation and by the MCMC approach of Carvalho (2006).

The remainder of this chapter is structured as follows: Section 4.2 discusses model identification for exploratory and confirmatory factor models and the statistical and numerical issues involved in these methods. Section 4.3 describes the MCMC approach for sparse factor analysis originally developed by Carvalho (2006) in its slightly altered version by Kaufmann and Schumacher (2013) and discusses how the approach can be affected by numerical issues. A short simulation study then investigates these numerical issues. Section 4.4 proposes an alternative two-step approach, Section 4.5 reports the results of a simulation study, where the method is applied to discover the correct number of factors, the set of relevant variables and the hidden parsimonious structure and is compared to the sampler of Kaufmann and Schumacher (2013). Section 4.6 analyzes the students test data by means of the novel approach in the same way with respect to the three questions of interest. Section 4.7 concludes.

4.2 Exploratory and Confirmatory Factor Analysis

Latent factor models reduce a large set of N observable variables to a potentially much smaller number K of unobservable factors. If variables over time are considered, a vector of N variables in time period t, denoted as y_t, where y_t is assumed to be demeaned, is thus represented as

$$y_t = \Lambda f_t + e_t, \quad \text{for all } t \in \{1, \dots, T\}, \tag{4.1}$$

where Λ is a $N \times K$ matrix of factor loadings, f_t is the $K \times 1$ vector of latent factors in time t, and e_t is a vector of zero-mean idiosyncratic error terms. Since the observed data is split up

into an explained part Λf_t and and unexplained part e_t, the factors and errors are assumed to be mutually uncorrelated, i.e.

$$\begin{pmatrix} f_t \\ e_t \end{pmatrix} \sim i.i.d. \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega & 0 \\ 0 & \Sigma \end{bmatrix} \right), \tag{4.2}$$

where $\Omega \in \mathbb{R}^{K \times K}$ and $\Sigma \in \mathbb{R}^{N \times N}$.

The earliest developed estimation procedure, Principal Components (PC) factor analysis, yields a unique estimate satisfying this specification. Factors are constructed from the scaled eigenvectors of the empirical covariance or correlation matrix. More precisely, the PC solution to the factor model in Equation (4.1) is

$$\{\hat{\Lambda}, \{\hat{f}_t\}_{t=1}^T\} = \underset{\Lambda, \{f_t\}_{t=1}^T}{\operatorname{argmin}} \left\{ \sum_{t=1}^T (y_t - \Lambda f_t)'(y_t - \Lambda f_t) \right\}, \tag{4.3}$$

subject to

$$\frac{1}{T} \sum_{t=1}^T f_t' f_t = I_K \quad \text{and} \quad \Lambda'\Lambda = \text{diag}. \tag{4.4}$$

Since the eigenvectors are orthogonal by definition, the same holds for the factors. Moreover, they are ordered by decreasing order of magnitude of their corresponding eigenvalues, so the solution is unique as long as the eigenvalues are. More details on PC estimation of static factors are found e.g. in Bai and Ng (2013).

To apply maximum likelihood (ML) or Bayesian factor analysis, distributional assumptions are needed. Usually, the factors and errors are assumed to be normally distributed, i.e. Equation (4.2) changes to

$$\begin{pmatrix} f_t \\ e_t \end{pmatrix} \sim f_N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega & 0 \\ 0 & \Sigma \end{bmatrix} \right), \tag{4.5}$$

and the likelihood function obtains as

$$\mathcal{L}\left(\{y_t\}_{t=1}^T | \{f_t\}_{t=1}^T, \Lambda, \Sigma \right) = \prod_{t=1}^T (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (y_t - \Lambda f_t)' \Sigma^{-1} (y_t - \Lambda f_t) \right). \tag{4.6}$$

If the factors are integrated out from Equation (4.6), the resulting marginalized likelihood is

$$\mathcal{L}(\{y_t\}_{t=1}^T | \Lambda, \Sigma, \Omega) = (2\pi)^{-\frac{TN}{2}} |\Lambda\Omega\Lambda' + \Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^T (y_t'(\Lambda\Omega\Lambda' + \Sigma)^{-1} y_t) \right\}. \tag{4.7}$$

Initially, factor analysis did not deal with time series data, but with cross-sectional data only. Instead of time points t, objects (or subjects) s are considered, for each of which N characteristics can be observed, which are governed by the $K << N$ underlying factors. The

concept of factor models evolved in psychometrics, with single factor analysis developed by Spearman (1904), and multiple factor analysis by Thurstone (1935). The models were mostly concerned with evaluating intellectual capabilities of subjects' test results in order to find out whether certain abilities could be attributed to one or multiple intelligence factors. The data set by Holzinger and Swineford (1939) used in the application section of this chapter likewise contains data obtained from a series of tests measuring intellectual capabilities of high school students.

There exist several extensions to the factor model in Equation (4.1), allowing for the introduction of dynamics, modeling the factors as vector autoregressive (VAR) processes or for serial correlation in the idiosyncratic error terms e_t. To keep things simple, I consider a static factor model without serial correlation in the error terms, and moreover, assume that there is no cross-correlation in the error terms, either, i.e. $\Sigma = \text{diag}(\sigma_1^2, \ \ldots, \ \sigma_N^2)$ is diagonal, corresponding to the *Frisch case* of Scherrer and Deistler (1998), named after the specification originally used by Frisch (1934). The empirical application likewise uses a cross-sectional data set without time dynamics. It must be pointed out, however, that the method introduced here is likewise suitable for the dynamic factor model. An application for the dynamic model, however, will be the subject of Chapter 5.

4.2.1 Exploratory Factor Analysis and Model Identification

In exploratory factor analysis (EFA), no assumptions about the structure of the loadings matrix Λ are made. Two exceptions apply, however, the first being the model indeterminacy that can be separated into a *scaling indeterminacy* and a *rotation indeterminacy* in the case of orthogonal factors, and the second being an upper bound to the number of factors K that can be estimated from the available information, known as the *Ledermann bound*.

Consider the model indeterminacy first: The model in Equation (4.1) is not uniquely identified, since both the factors $\{f_t\}_{t=1}^T$ and the factor loadings matrix Λ are unobservable, and hence Equation (4.1) can be written as

$$y_t = \underbrace{\Lambda D}_{\Lambda^*} \underbrace{D^{-1} f_t}_{f_t^*} + e_t, \quad \text{for all } t \in \{1, \ldots, T\}, \tag{4.8}$$

where D is a $K \times K$ invertible matrix, and Λ^* and $\{f_t^*\}_{t=1}^T$ are a set of alternative factor loadings matrices and factors, whereas e_t remains unchanged. The model parameter Σ therefore likewise remains unchanged, whereas Ω becomes $\Omega^* = D^{-1}\Omega D^{-1\prime}$. The matrix D implies that K^2 model parameters must be fixed in a suitable way to identify the model. If an orthogonal factor model is considered, i.e. the factors $\{f_t\}_{t=1}^T$ are uncorrelated, Ω is a diagonal matrix. This fixes $\frac{K(K-1)}{2}$ parameters. Next, it is possible to scale each $N \times 1$ column of the loadings matrix Λ by some positive value and the corresponding factor by its inverse. To prevent this, the variance of the factors is fixed to unity, hence fixing another K parameters up to their sign. The remaining $\frac{K(K-1)}{2}$ free parameters now constitute the

well-known rotation problem, see e.g. Thurstone (1935) or Anderson and Rubin (1956). The rotation problem can be partially fixed by imposing a lower triangular (LT) structure on the loadings matrix. As a result, the model is identified up to the signs of the factors and the corresponding loadings. A QR decomposition of the transpose of Λ into

$$\Lambda' = QR = D'\Lambda'_{LT} \tag{4.9}$$

yields an upper triangular matrix R and an orthogonal matrix Q. Postmultiplication of Λ by the orthogonal matrix $D = Q'$ therefore results in the lower triangular matrix $\Lambda_{LT} = R'$. As Λ remains lower triangular even under a reflection of any subset of factors and corresponding columns of Λ, the model is *locally identified* in this case, see Anderson and Rubin (1956).[3]

Equation (4.9) can be expanded by a reflection matrix B, i.e. a diagonal matrix whose diagonal elements are either 1 or -1, which yields

$$\Lambda' = \underbrace{QB}_{D'} \underbrace{B'R}_{\Lambda'_{PLT}}. \tag{4.10}$$

Since B is an orthogonal matrix and the product of two orthogonal matrices is an orthogonal matrix, and the inverse of an orthogonal matrix is also orthogonal, D is a unique orthogonal matrix and the matrix Λ_{PLT} is positive lower triangular (PLT) and the unique positive lower triangular matrix that can be obtained from Λ by an orthogonal transformation. Hence, demanding positive lower triangularity for Λ solves the rotation problem and the model is *globally identified*. For a rigorous proof, see e.g. Muirhead (1982), Theorem A9.8.

If the top $K \times K$ submatrix of Λ, denoted as Λ_a, has a rank deficit, $K - \text{rk}(\Lambda_a)$ diagonal elements of Λ_{PLT} become zero in the above transformation, hence the PLT constraint cannot be satisfied and the rotation problem cannot be solved. Moreover, rearranging the rows of Λ in such a way that at least one out of the first K rows ends up in a different place, i.e. a change of Λ_a, results in a different matrix Λ_{PLT}.[4] Such a rearrangement can be performed using an $N \times N$ permutation matrix O, and the model from Equation (4.1) accordingly becomes

$$Oy_t = O\Lambda f_t + Oe_t, \quad \text{for all } t \in \{1, \ldots, T\}, \tag{4.11}$$

so Σ becomes $O\Sigma O'$. The rearrangement leaves the factors unchanged and only permutes the rows of Λ and $Y = (y_1, \ldots, y_T)'$ and the rows and columns of Σ, so the model is still the same.[5] As the model identification hinges on Λ_a, and thus on the ordering of Y, the PLT constraint may be unattainable for some choices of O. Due to their importance for model identification, the variables placed in the first K rows of Y are called the *factor founders* by Carvalho et al. (2008). A particular choice of the factor founders leads to a particular shape

[3] Local identification allows for the existence of multiple identical maxima of the likelihood, which have to be clearly separated from each other, however.

[4] Unless, of course, a variable is replaced by another with identical values throughout.

[5] In the Frisch case considered here, only the diagonal elements of Σ are permuted, as all off-diagonal elements of Σ are zero anyway.

of the PLT constrained likelihood, which consists of all points satisfying the PLT constraint imposed on Λ for identification.

Now consider the Ledermann bound, which constrains the number of factors, and accordingly, the number of columns of Λ. In the case where $N < T$, the $N \times N$ covariance matrix of Y has full rank. In order to obtain a unique separation between the systematic part $\{\Lambda f_t\}_{t=1}^T$ and the idiosyncratic part $\{e_t\}_{t=1}^T$ in Equation (4.1), the number of factors K must not exceed the Ledermann bound, see Ledermann (1937). This bound is derived from counting the number of different elements in the covariance matrix, which is $\frac{N(N+1)}{2}$, or in the correlation matrix, which is $\frac{N(N-1)}{2}$. The number of parameters to be estimated is then $NK - \frac{K(K-1)}{2} + N$ or $NK - \frac{K(K-1)}{2}$, where Λ contains NK parameters, the factors are mutually orthogonal and have variance one each. In the case of the correlation matrix, Λ automatically determines the nonzero elements of Σ, so the σ_i^2 are no longer free parameters. The *rotation problem* described below introduces the matrix D, which reduces the number of uniquely determined parameters by $\frac{K(K-1)}{2}$. Solving for K, the Ledermann bound obtains as

$$\varphi(N) = \frac{2N + 1 - \sqrt{8N+1}}{2}. \tag{4.12}$$

The assumption that $N < T$ is not in line with the "large p small n" paradigm central to sparse factor analysis, see West (2003), so taking the reduced rank of the covariance matrix of Y in such a case into account, there remain only $\frac{T(T+1)}{2}$ distinct elements, and the Ledermann bound accordingly changes to

$$\varphi(N,T) = \frac{2N + 1 - \sqrt{4(N+T)(N-T+1)+1}}{2} \quad \text{for} \quad T \leq N, \tag{4.13}$$

which nests the expression in Equation (4.12) for $T = N$.

4.2.2 Numerical Issues in Exploratory Factor Analysis

Numerical issues in factor analysis may arise due to the shape of the likelihood that results from imposing the necessary identification constraints. This can be illustrated by looking at the identification via an LT constraint on Λ first, such that 2^K modes of the likelihood for different choices of Λ exist. Starting from one mode, switching a subset of the column signs of Λ leads to a new mode with the same likelihood value. Consider a model with $K = 2$ factors. There are hence $2^2 = 4$ identical modes of the likelihood under a particular LT constraint. The four reflection matrices performing the switching are then

$$D_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_K, \quad D_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad D_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{4.14}$$

$$\text{and} \quad D_4 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \cos(\pi) & -\sin(\pi) \\ \sin(\pi) & \cos(\pi) \end{pmatrix}, \tag{4.15}$$

where D_1 is the identity matrix, D_2 and D_3 are matrices with negative determinants, and D_4 is a rotation matrix around the angle $\gamma = \pi$.

If Λ is postmultiplied by any of the four orthogonal matrices D_1, D_2, D_3 and D_4, the likelihood function in Equation (4.7) with the scaling constraint $\Omega = I_K$ yields the same value, because Λ only enters this expression in the form of its outer product, where the effect of the orthogonal matrices cancels out. To illustrate the effect of the LT constraint on the likelihood surface, Loken (2005) uses a graphical illustration, which is slightly extended and discussed in the following. Instead of the four matrices discussed above, consider arbitrary orthogonal matrices D postmultiplied to Λ. These matrices D may be rotation matrices, which have determinant 1 and are expressible for $K = 2$ in terms of a single parameter γ, or reflection matrices with determinant -1, which are expressible for $K = 2$ as the product of a permutation matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and a rotation matrix with the aforementioned properties. For arbitrary orthogonal matrices D, $\Lambda^\dagger = \Lambda D$ generally does not satisfy the LT constraint. Therefore, $\lambda^\dagger_{1,2}$ is set to zero and the resulting expression is used to evaluate the likelihood.

Figure 4.1 shows the log likelihood as a function of the angle γ as a solid line, as in Loken (2005). Along this solid line, two of the aforementioned modes can be seen, the first implied by $\gamma = 0$, which produces the matrix D_1, and the second implied by $\gamma = \pi$, which produces the matrix D_4. The dashed line shows the log likelihood as a function of γ if the columns of Λ are exchanged before the rotation, implying that $\lambda^\dagger_{1,1}$ is set to zero. The remaining two modes can be seen along this line: The product of P and a rotation matrix with $\gamma = -\frac{\pi}{2}$ produces the matrix D_2 and the product of P and a rotation matrix with $\gamma = \frac{\pi}{2}$ produces the matrix D_3. Note, however, that the constraint used in Figure 4.1 is not the only admissible one. If the variables in Y are rearranged and the LT constraint is applied to the correspondingly rearranged Λ in the same way as before, there are still four modes, still related to each other by the four matrices D_1 through D_4. Instead of rearranging the variables, it is possible to simply rearrange the constraints. Hence exchanging e.g. variables 1 and 4 and imposing the LT constraint on the orthogonally transformed Λ and setting $\lambda^\dagger_{1,2}$ to zero has the exact same effect as leaving the variables in their original ordering and setting $\lambda^\dagger_{4,2}$ to zero. Placing the zero constraint on a different element of $\Lambda^\dagger$ amounts to replacing one local identification of the model by another local identification.

For general K, the LT constraint requires that elements in the first $K-1$ rows of Λ be set to zero. Accordingly, the first $K-1$ variables in Y serve as factor founders. Note that the ordering among these $K-1$ variables also matters, as e.g. the loadings of the first variable on all factors except for the first are set to zero, whereas for the $(K-1)^{\text{th}}$ variable, only the loading on the last factor is set to zero. Thus there are $\binom{N}{K-1}(K-1) = \frac{N!}{(N-K+1)!}$ different ways to locally identify the model using the LT identification constraint. Since each of them identifies the system in an equivalent manner, the maxima of the log likelihood function must be identical. Outside the maxima, however, the behavior of the log likelihood as a function of the angle γ is different. For the aforementioned example, which requires only a single factor

founder for the LT constraint, the case where $\lambda_{4,2}^{\dagger}$ is constrained to zero is shown in Figure 4.2, superimposed over the initial case where $\lambda_{1,2}^{\dagger}$ is constrained to zero.[6] Figure 4.2 shows that the maxima of the log likelihood under the alternative constraint are the same, but the minima are much larger than the minima under the initial constraint. Table 4.1 shows the minima of the constrained log likelihood for the $N = 10$ different choices of the factor founder as well as the location of the mode in the first quadrant.[7] These minima substantially differ from each other, implying that the choice of the factor founder has a strong impact on the shape of the constrained likelihood. As it may be difficult to find the mode of a rather flat likelihood, the factor founders should be chosen such that this is ruled out, or, equivalently, the variables should be ordered accordingly. If instead of the LT constraint, the PLT constraint is used, K instead of $K-1$ factor founders must be chosen and inequality constraints are imposed on the diagonal elements of Λ, which cause the constrained likelihood to lose its symmetric shape. Breaking the symmetry is in fact the intended consequence of imposing the PLT constraint, such that only a single global mode remains. As a side effect, however, the likelihood surface may have nearly flat regions as well as local modes. This case is analyzed in Chapter 3.

In Bayesian factor analysis, where an effect of the ordering of the variables on the shape of the posterior distribution has been observed e.g. by Lopes and West (2004) or Frühwirth-Schnatter and Lopes (2012), this issue is therefore referred to as the *ordering problem*. Bayesian factor analysis generally uses the PLT constraint, which results in a highly asymmetric posterior distribution with flat stretches and potential local modes even if an orthogonally invariant prior distribution is used.

4.2.3 Confirmatory Factor Analysis and Model Identification

Confirmatory factor analysis (CFA), as opposed to EFA, introduces assumptions about the association between factors and variables and hence about the structure of Λ via fixed zero elements in the loadings matrix.[8] The number of zero elements in Λ exceeds the required number of zero elements to ensure exact local or global identification, as in the case of the LT constraint discussed in Section 4.2.1. Whereas in CFA a substantial share of the elements of the loadings matrix may be set to zero, the orthogonality assumption about the factors is often relaxed, such that the off-diagonal elements of Ω are different from zero, see e.g. Rubin and Thayer (1982) or Mulaik (2010). The imposed structure of the loadings matrix is derived from a-priori hypothetical reasoning, and CFA allows for testing these hypotheses based on the outcome of the constrained analysis. Tests for the structure of Λ as well as goodness-of-fit measures, see e.g. Jöreskog and Sörbom (1986) and Bentler (1990), include those used in structural equation modeling (SEM), see e.g. Bollen (1989), are available. In fact, CFA is a special case of SEM, see Mulaik (2010).

[6] The slightly shifted modes of the alternative constraint indicate that starting from the initial Λ, which has $\lambda_{1,2} = 0$, a slight rotation to the left is necessary in order to have $\lambda_{4,2} = 0$.

[7] Identical modes exist in the remaining three quadrants at interval $\frac{\pi}{2}$, see above or Loken (2005).

[8] A less common approach in CFA is to fix some elements of Λ to values other than zero or to use inequality constraints.

Some frequently used predefined loadings structures in CFA are the congeneric factor model, see Jöreskog (1971), also known as dedicated factor model, see Conti et al. (2014), which allows each variable to load on exactly one factor, and the aforementioned bi-factor model of Holzinger and Swineford (1937, 1939), in which there is a common factor that all variables load on and each variable loads on one additional group-specific factor. Many other loadings structures are possible, where the imposed structure for Λ in CFA generally nests an exact model identification such as the LT constraint, so the rotation problem is generally not an issue any more. In some cases, however, adding zero constraints to Λ on top of the LT constraint results in the model being not identified any more.

Howe (1955) and Jöreskog (1969) give conditions for the uniqueness of the solution under a pre-specified loadings structure. Dunn (1973) and Jennrich (1978), however, provide examples where these conditions are not sufficient. The example by Dunn (1973) starts from the LT identification constraint, hence the top $K \times K$ section of Λ is

$$\Lambda_a = \begin{pmatrix} \lambda_{1,1} & 0 & 0 \\ \lambda_{2,1} & \lambda_{2,2} & 0 \\ \lambda_{3,1} & \lambda_{3,2} & \lambda_{3,3} \end{pmatrix}. \tag{4.16}$$

If an additional zero constraint is imposed on $\lambda_{2,2}$, the matrix in (4.16) becomes

$$\Lambda_a = \begin{pmatrix} \lambda_{1,1} & 0 & 0 \\ \lambda_{2,1} & 0 & 0 \\ \lambda_{3,1} & \lambda_{3,2} & \lambda_{3,3} \end{pmatrix}. \tag{4.17}$$

The local model identification is lost, as a postmultiplication of the matrix in (4.17) by a rotation matrix

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\gamma) & -\sin(\gamma) \\ 0 & \sin(\gamma) & \cos(\gamma) \end{pmatrix} \tag{4.18}$$

with arbitrary angle γ still satisfies the zero constraints. The example by Jennrich (1978) shows that fixing all $N(K-1)$ elements in the first $K-1$ rows of Λ does not guarantee uniqueness of Λ, since a reflection about the space spanned by these rows allows for a second solution with identical elements in the first $K-1$ rows, but different elements in the remaining rows. This fact, also noted by Geweke and Singleton (1981), leads to the conclusion that a PLT structure in the loadings matrix is a necessary and sufficient condition for model identification, and any sparse structure that nests a PLT structure is therefore also identified. Bekker (1986) provides generalized necessary and sufficient conditions for identification in CFA.

The number of sets of constraints that ensures model identification is generally very large. Hence an important question, raised e.g. by Millsap (2001), concerns the placement of the overidentifying constraints in Λ. To find appropriate placements for the zero elements in Λ, an EFA can be done first, whose result can then be orthogonally or obliquely transformed,

depending on whether the factors are supposed to be orthogonal or may be oblique. Such transformations to discover a *simple structure* or *simple configuration*, i.e. a loadings matrix with many elements close to zero, proposed by Thurstone (1938), may generally facilitate interpretability of the results, but may also indicate which elements of Λ can be constrained to zero in a subsequent CFA. To find a simple structure, several optimization criteria can be used. A popular choice is the Varimax criterion by Kaiser (1958), which attempts to find a rotation of the initial PC factor solution in which each variable is driven predominantly by a single factor, and, accordingly, the factors are determined by as few variables as possible.

Which elements of Λ the constraints are placed on has both statistical and numerical implications. On the statistical side, goodness-of-fit criteria can be applied to compare different model setups, see e.g. Marsh et al. (1988) or Cheung and Rensvold (2002). This requires, however, that estimates are obtained in the first place, which may become difficult for numerical reasons discussed in the following subsection.

4.2.4 Numerical Issues in Confirmatory Factor Analysis

Confirmatory factor analysis faces numerical issues beyond those described in Section 4.2.2 for EFA. Under a fixed structure of Λ, the surface of the likelihood may have multiple modes and generally have a less regular shape than in EFA with minimal global identification constraints, such as the PLT constraint. Rubin and Thayer (1982) apply the expectation-maximization (EM) algorithm in ML factor analysis for exploratory and confirmatory factor analysis. Analyzing a data set from Jöreskog (1969) by CFA, they report finding multiple maxima. Bentler and Tanaka (1983) refute this claim and attribute the result to convergence problems of the EM algorithm. In a reply, Rubin and Thayer (1983) note that the EM algorithm in fact tends to converge to different maxima in high-dimensional problems such as the considered example.

The multimodality issue observed here can be illustrated by an example similar to the one discussed in Section 4.2.2 for EFA. Again, a model setup with $K = 2$ is considered, Λ is transformed by postmultiplying an orthogonal matrix D, which may be a rotation matrix with determinant 1, expressible in terms of a single parameter γ, the rotation angle, or a reflection matrices with determinant -1, expressible as the product of the permutation matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and a rotation matrix with the aforementioned properties. The initial zero constraints are again imposed on the resulting matrix $\Lambda^\dagger = \Lambda D$, based on which the log likelihood is evaluated. Instead of the LT constraint, 45% of the elements of Λ are set to zero. Figure 4.3 shows the outcome of the experiment for two different structures of Λ, where the second is a permutation of the first. It can be seen that the global maxima are still related to each other by the four matrices D_1 through D_4, i.e. there still exists one global maximum per quadrant, whereas the minima are no longer located exactly in the middle between two maxima. Erosheva and Curtis (2013) argue that in Bayesian CFA using the Gibbs sampler, a global identification constraint is not required and may even obstruct the sampler, an effect

similar to that of the PLT constraint in exploratory factor analysis described in Chapter 3. Instead, the Gibbs output can be postprocessed with a relabeling algorithm similar to that of Stephens (2000). In the example with $K = 2$, this would imply mapping draws from the remaining three quadrants into the first quadrant.

If the rows in the initial structure of Λ undergo a permutation, additional local maxima evolve as a result of the overidentifying constraints. Table 4.2 shows the results from ten random row permutations, each implying a different structure of zero constraints on Λ, and hence a different model representation. Two out of the ten model representations allow for local maxima in addition to the global maxima. The log likelihood value in the maximum as well as the locations of the maximum in the first quadrant expressed in terms of the rotation angle γ are also given in the table. The reported maxima may serve to evaluate the model fit: The first permutation uses the set of constraints used for generating the data and reaches the highest log likelihood value. The second permutation, however, reaches almost the same log likelihood value in the maximum, but a much lower log likelihood value in the minimum, so the shapes of the constrained likelihoods under the first two models differ substantially.

4.3 Sparse Factor Analysis

Sparse factor analysis is an exploratory technique that yields estimates for Λ that have the same properties as estimates from CFA, but it does not require that the structure in Λ is postulated prior to the analysis. Instead, the zero elements in Λ are identified by the approach, which can hence be described as "self-organizing".[9] In terms of Bayesian sparse factor analysis, which this section focuses on, CFA can therefore be understood as a special case, where the association probabilities between the factors and variables, i.e. the probabilities of the elements of Λ for being equal to zero, are all set either to zero or to one.

Model identification is generally not an issue in sparse factor analysis, as the number and location of the zero elements in Λ generally nests the LT constraint. If the number of zero elements gets very large, however, the three-indicator rule for confirmatory factor analysis with uncorrelated factors must be considered, see e.g. Bollen (1989), p.274. This rule builds on Theorem 5.5 from Anderson and Rubin (1956) stating a sufficient condition to identify a factor, which is that three of its loadings must be different from zero.[10]

Regarding frequentist approaches, an estimation procedure to generate a sparse PC factor representation based on Lagrange multipliers has been suggested by Charles (1998). Similarly, penalized least-squares methods, such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1994) and its extension, the elastic net (EN) (Zou and Hastie, 2005), have been used by Zou et al. (2006) to find a sparse PC representation. Both approaches have been implemented in Bayesian analysis, see e.g. Park and Casella (2008) or Li and Lin (2010), while Bayesian sparse factor analysis as considered in the following builds on the spike and

[9] Mike West used this expression in a presentation in 2014 to characterize Bayesian sparse factor analysis.
[10] For correlated factors, an according two-indicator rule exists.

slab prior by Mitchell and Beauchamp (1988), which found many applications for variable selection. For a review, see e.g. Ishwaran and Rao (2005).

4.3.1 Bayesian Sparse Factor Analysis

Bayesian sparse factor analysis was first proposed by West (2003). The core element of this approach is the sparse prior for Λ, which is a mixture of a Dirac delta function δ_0 and a normal prior, where the prior probability of an element of the loadings matrix to be zero is in turn governed by β_k, which follows a beta distribution. Hence, the prior for every $\lambda_{i,k}$ with $i \in \{1, \ldots, N\}$ and $k \in \{1, \ldots, K\}$ is e.g.

$$\lambda_{i,k} \sim \beta_k \delta_0(\lambda_{i,k}) + (1 - \beta_k) f_N(\lambda_{i,k} | 0, \tau_k) \tag{4.19}$$

with association probabilities $1 - \beta_k$, where

$$\beta_k \sim f_B(\beta_k | s_k r_k, s_k(1 - r_k)). \tag{4.20}$$

This prior is the one-layer sparse prior, where e.g. for $s_k = 1000$ and $r_k = 0.999$, the a-priori share of zeros is 99.9% for the loadings of factor k, hence the prior association probability $1 - \beta_k$ of associations between all variables and factor k is 0.1%, and τ_k governs the prior variance of the non-zero loadings on factor k. The prior association probability β_k can be varied across the factors, but stays the same for all loadings of the same factor.

The two-layer sparse prior, proposed by Lucas et al. (2006), changes Equations (4.19) and (4.20) to

$$\lambda_{i,k} \sim (1 - \beta_{i,k}) \delta_0(\lambda_{i,k}) + \beta_{i,k} f_N(\lambda_{i,k} | 0, \tau_k) \tag{4.21}$$

$$\beta_{i,k} \sim (1 - \rho_k) \delta_0(\beta_{i,k}) + \rho_k f_B(\beta_{i,k} | s_k r_k, s_k(1 - r_k)) \tag{4.22}$$

$$\rho_k \sim f_B(\rho_k | w_k v_k, w_k(1 - v_k)) \tag{4.23}$$

so the $\beta_{i,k}$ are association probabilities that are individual for each element of the loadings matrix, and ρ_k is the base rate of the non-zero factor loadings. Based on this flexible modeling, Bhattacharya and Dunson (2011) propose an approach to shrink the loadings of factors to zero as k increases, and hence allows to choose the number of factors required in the model accordingly.

4.3.2 A Sampler for Sparse Factor Analysis

Frühwirth-Schnatter and Lopes (2012) suggest a sampling scheme that is not only able to identify a sparse structure, but also to determine the number of factors. They apply their scheme to several data sets, including simulated data, where they find that with the proper prior specification, the simulated structure is recovered very well by the model. Carvalho (2006) proposes a sampler that uses a two-layer sparse prior, which is slightly altered by

Kaufmann and Schumacher (2012, 2013). In a simulation study, the sampler is shown to perform best for high degrees of sparsity and to identify irrelevant variables, i.e. variables consisting only of noise and hence having zero loadings of every factor. Kaufmann and Schumacher (2012) find that the minimal identification constraints in terms of the (P)LT constraint imposed by Carvalho (2006) are generally not required, as they are nested in the sparse structure identified by the sampler. They use an alternative identification constraint to deal with the rotation problem, however.

To generate a sample from the posterior distribution of the parameters and the factors, the sampler iteratively draws from the full conditional distributions of Λ, Σ and $F = (f_1, \ldots, f_T)'$ and thus resembles the Gibbs sampler used for the static factor model introduced in Chapter 2 and used in the proposed two-step approach. This Gibbs sampler is summarized in Appendix 4.A. While the sampling steps for the idiosyncratic variances Σ and the factors F are identical, the sampling step for the sparse loadings matrix Λ is more complicated. In the following, the sampling approach by Carvalho (2006), as described in Kaufmann and Schumacher (2013) is used. Unlike Carvalho (2006), Kaufmann and Schumacher (2013) do not apply the PLT identification constraint, but an identification scheme suggested by Anderson and Rubin (1956), demanding that $\Lambda'\Lambda$ is diagonal with its diagonal elements in descending order of magnitude. This provides a local identification in the same way as the LT scheme, so in a subsequent step, the signs of the factors and the columns of Λ have to be readjusted. Kaufmann and Schumacher (2013) achieve this by switching the sign if the correlation with the remaining draws is negative. The identification scheme is applied by temporarily transforming $\Lambda^{(z)}$ to $\Lambda_*^{(z)}$ such that $\Lambda_*^{(z)'}\Lambda_*^{(z)}$ satisfies the aforementioned diagonality constraint. The matrix of factors is then sampled as $F_*^{(z)}$ conditional on $\Lambda_*^{(z)}$, and to obtain $F^{(z)}$, the reverse of the initial orthogonal transformation is applied to $F_*^{(z)}$.

Obtaining the posterior distribution of the elements of the sparse loadings matrix $\lambda_{i,k}$ requires to first integrate out the probability of a zero loading for factor k on variable i, and hence to obtain the marginal prior distribution from the two-layer prior in Equation (4.22). The marginal prior is then

$$\pi(\lambda_{i,k}|\rho_k) = (1 - \rho_k r_k)\delta_0(\lambda_{i,k}) + \rho_k r_k f_N(\lambda_{i,k}|0, \tau_k). \tag{4.24}$$

In order to only consider the relation between variable i and factor k, the effect of the remaining factors on the variable is removed, which yields

$$\tilde{y}_{i,t,k} = y_{i,t} - \sum_{j=1, j\neq k}^{K} \lambda_{i,j} f_{t,j} = \lambda_{i,k} f_{k,t} + e_{i,t}, \tag{4.25}$$

so

$$\tilde{y}_{i,t,k} \sim f_N(\lambda_{i,k} f_{t,k}, \sigma_i^2). \tag{4.26}$$

The full conditional distribution of $\lambda_{i,k}$ is then

$$g(\lambda_{i,k}|\cdot) = \prod_{t=1}^{T} f(\tilde{y}_{i,t,k}) \left\{(1-\rho_k r_k)\delta_0(\lambda_{i,k}) + \rho_k r_k f_N(\lambda_{i,k}|0,\tau_k)\right\} \tag{4.27}$$

$$= p(\lambda_{i,k}=0)\delta_0(\lambda_{i,k}) + p(\lambda_{i,k}\neq 0) f_N(\lambda_{i,k}|m_{i,k},M_{i,k}), \tag{4.28}$$

where $M_{i,k} = \left(\sigma_i^{-2}\sum_{t=1}^{T} f_{t,k}^2 + \tau_k^{-1}\right)^{-1}$ and $m_{i,k} = M_{i,k}\left(\sigma_i^{-2}\sum_{t=1}^{T} f_{t,k}\tilde{y}_{i,t,k}\right)$.

Calculating the posterior odds of variable i having a nonzero loading on factor k requires updating the prior odds that can be obtained from Equation (4.24), and yields

$$\frac{p(\lambda_{i,k}\neq 0|\cdot)}{p(\lambda_{i,k}=0|\cdot)} = \frac{\rho_k r_k}{1-\rho_k r_k}\cdot\frac{f_N(0|0,\tau_k)}{f_N(0|m_{i,k},M_{i,k})}. \tag{4.29}$$

Next, the association probability $\beta_{i,k}$ for each $\lambda_{i,k}$ is sampled from its full conditional distribution, where for $\lambda_{i,k}=0$,

$$g(\beta_{i,k}|\lambda_{i,k}=0,\cdot) \propto (1-\beta_{i,k})\left\{(1-\rho_k)\delta_0(\beta_{i,k}) + \rho_k f_B(s_k r_k, s_k(1-r_k))\right\}, \tag{4.30}$$

where

$$p(\beta_{i,k}=0|\lambda_{i,k}=0,\cdot)\propto 1-\rho_k \qquad \text{and} \quad p(\beta_{i,k}\neq 0|\lambda_{i,k}=0,\cdot)\propto (1-r_k)\rho_k. \tag{4.31}$$

Therefore, if $\lambda_{i,k}=0$, the posterior odds for sampling $\beta_{i,k}$ from $f_B(s_k r_k, s_k(1-r_k)+1)$ versus $\beta_{i,k}=0$ are $\frac{(1-r_k)\rho_k}{1-\rho_k}$.

Conversely, for $\lambda_{i,k}\neq 0$,

$$g(\beta_{i,k}|\lambda_{i,k}\neq 0,\cdot) \propto \beta_{i,k} f_N(\lambda_{i,k}|0,\tau_k)\left\{(1-\rho_k)\delta_0(\beta_{i,k}) + \rho_k f_B(s_k r_k, s_k(1-r_k))\right\}, \tag{4.32}$$

where

$$p(\beta_{i,k}=0|\lambda_{i,k}\neq 0,\cdot)=0 \qquad \text{and} \quad p(\beta_{i,k}\neq 0|\lambda_{i,k}\neq 0,\cdot)=1. \tag{4.33}$$

Therefore, if $\lambda_{i,k}\neq 0$, $\beta_{i,k}$ is sampled from $f_B(s_k r_k+1, s_k(1-r_k))$.

The hyperparameters τ_k and ρ_k are then updated by sampling from

$$g(\tau_k|\cdot) = f_{IG}\left(c_0 + \frac{1}{2}\sum_{i=1}^{N} I_{\{\lambda_{i,k}\neq 0\}}, C_0 + \frac{1}{2}\sum_{i=1}^{N}\lambda_{i,k}^2\right) \tag{4.34}$$

and

$$g(\rho_k|\cdot) = f_B\left(w_k v_k + \sum_{i=1}^{N} I_{\{\beta_{i,k}\neq 0\}}, w_k(1-v_k) + N + \sum_{i=1}^{N} I_{\{\beta_{i,k}\neq 0\}}\right). \tag{4.35}$$

In order to account for the posterior density of $\lambda_{i,k}$ being a two-component mixture, the estimator for the sparse loadings matrix Λ is not the mean of the obtained sequences for the $\lambda_{i,k}$, but the median.

4.3.3 Some Numerical Issues in Sparse Factor Analysis

The numerical issues in sparse factor analysis are related to those in CFA, but the sampler's flexibility adds to their complexity. The example in Section 4.2.4 illustrates that the constrained likelihood in CFA may have additional local modes. Accordingly, the posterior distribution of Λ under a specific structure may have local modes, which obstruct the sampler's mixing behavior, which may hence stay in the vicinity of such a local mode, failing to reach the global mode. The structure in Λ, however, may change in the process of sampling, so if the degree of sparsity is fixed, there exists a variety of different local modes. Consider the example from Section 4.2.4, where ten row permutations of the zero constraints yield the corresponding maxima of the log likelihood given in Table 4.2. All these maxima are modes of the log likelihood in a sparse model with fixed degree of sparsity. Additional modes can be found by considering other permutations of the zero constraints, or by allowing the degree of sparsity to vary.

The sampler may, however, never visit all these modes. If sufficiently many elements of Λ are shrunk to zero quickly, the number of zero constraints is sufficiently large such that the constraints normally imposed to solve the rotation problem are no longer necessary. The behavior of different sparsity priors in this respect has been analyzed by Malsiner-Walli and Wagner (2011), who point out that sparsity priors are not well able to discriminate between small and zero effects. Hence there may be a danger of the sampler converging quickly to one mode of the posterior distribution, which, if spike and slab priors are used, has been observed to generate multimodal posterior densities, see e.g. Titsias and Lázaro-Gredilla (2011) and Ma and Zhao (2013). With the sampler stuck in a local mode, it may be impossible to reach other local modes or the global mode. Aside from this issue, the results of sparse factor analysis may be sensitive to the choice of the hyperparameters (Frühwirth-Schnatter and Lopes, 2012) and its performance may depend on the degree of sparsity in the data, see Kaufmann and Schumacher (2013).

4.3.4 Exploring Multimodality in Sparse Factor Analysis

A small simulation study may shed light on whether multiple solutions can be found. Four different loadings matrices are used to simulate one data set each. Each of them has $N = 50$ rows and $K = 3$ columns. The first loadings matrix contains 73% zero elements, the second contains 55% zero elements, the third contains 34% zero elements, and the fourth contains no zero elements, but 51% of the elements are less than 0.01 in magnitude, and 67% are less than 0.05 in magnitude. As Kaufmann and Schumacher (2013) find that their approach works best for large degrees of sparsity, the algorithm should be able to recover the sparse

loadings matrix well for the first data set. The algorithm also yields a sparse loadings matrix for the fourth data set, even though there is no exact sparse structure in it, but rather an approximate sparse structure. Empirically, such a case may occur as a result of measurement errors in the variables, or simply if the data does not support an exact sparse representation.

To better explore the space of sparse Λ matrices, I deviate from the common practice of initializing Λ with the Varimax rotation of its PC estimate. Instead, I perform 25 random orthogonal transformations of this point to use them as starting points for the sampler. Following Kaufmann and Schumacher (2013), the hyperparameters are chosen as $\underline{\alpha}_i = 1$ and $\underline{\beta}_i = 1$ for all $i \in \{1, \ldots, N\}$, $c_0 = 2$ and $C_0 = 0.5$, and $s_k = 1$, $r_k = 0.8$, $w_k = 500$ and $v_k = 0.1$ for all $k \in \{1, \ldots, K\}$. Every sequence consists of 15,000 iterations of the sampler, the first 5,000 of which are discarded as burn-in. Figure 4.4 shows the first 2,000 iterations of the sampler for the loadings on one randomly selected variable, i.e. one row of Λ in the scenario with 73% sparsity in Λ, where each replication uses a different starting point for Λ. For all ten starting points shown in the figure, the sampler quickly converges to the posterior distribution of the same sparse loadings matrix Λ. Next, the sum of squared loadings per factor is calculated from the posterior estimates of Λ, $\hat{\Lambda}$, for all 25 starting points. The results are shown in three scatter plots, one for each pair of factors, in Figure 4.8. The results are all very similar, so there is no evidence of an additional sparse representation.

Next, the same analysis is repeated for the data generated with 55% sparsity in Λ. Ten sequences of 15,000 iterations for one row of Λ, each obtained under different starting points, are shown in Figure 4.5. In most cases, convergence to the same distribution can be observed, even though this sometimes takes several thousand iterations of the sampler, as in replication 10. In replication 7 and 9, however, the 15,000 iterations do not suffice to converge to the distribution that the sampler converges to in the remaining replications. The sum of squared loadings per factor calculated from the posterior estimates, shown in the scatter plots in Figure 4.9, shows a large cluster and three outliers, which may imply that in these replications, the sampler failed to converge.

Afterwards, the analysis is repeated for the data generated with 34% sparsity in Λ. Ten sequences of 15,000 iterations for one row of Λ, each obtained under different starting points, are shown in Figure 4.6. While the factor loading with the sequence plotted in blue looks the same throughout the ten sequences, the sequences plotted in red and green show two distinct patterns, with the first prevalent in replications 1, 3, 9 and 10, and the second prevalent in the remaining replications. The scatter plots in Figure 4.10 even indicate the presence of three different sparse representations for Λ.

Eventually, the analysis is repeated for the data generated with the approximately sparse Λ. Ten sequences of 15,000 iterations for one row of Λ, each obtained under different starting points, are shown in Figure 4.7. The factor loading with the sequence plotted in blue apparently has two modes, where the sampler converges to the first in replication 1, 5, 6, 8 and 9, and to the second in replication 2, 3 and 4. In replication 7 and 10, the sampler

switches between both modes.[11] The factor loadings with the sequences plotted in green and red apparently have at least two modes, where the first is reached in replication 1 and 9, and the second is reached in the other replications. Accordingly, the sum of squared loadings per factor calculated from the posterior estimates, shown in the scatter plots in Figure 4.11, indicates that multiple sparse representations for Λ can be found.

This small simulation confirms the finding of Kaufmann and Schumacher (2013) that for a large degree of sparsity, the proposed sampler shows a good performance. In the scenarios with less sparsity, there may be multiple sparse representations for Λ and the estimate $\hat{\Lambda}$ appears to depend on the starting point of the sampler. The same holds for the scenario where the sparse pattern of Λ is artificially contaminated, representing either measurement errors, or the existence of merely an approximate sparse pattern in Λ. This finding may be especially relevant for the analysis of empirical data, in which a robustness check may involve starting the sampler from different points. An extended study may be useful to investigate whether changing the prior hyperparameters has an effect on the number of modes accessed by the sampler. Moreover, it may be worthwhile to find out if information from the sampled sequences can be used to decide whether additional sparse representations are likely to exist.

4.4 A Two-Step Approach

In this section, I propose a two-step approach whose first step consists of applying the weighted orthogonal Procrustes (WOP) estimation procedure described in Chapter 3. To avoid the numerical issues in exploratory factor analysis discussed in Section 4.2.2 and Chapter 3, in particular the ordering problem, a Gibbs sampler in which no (P)LT constraint is imposed on Λ is run. The output of this sampler is orthogonally mixed, i.e. the rotation problem must be solved in a postprocessing step, which is achieved by the WOP algorithm. The accordingly postprocessed output can then be considered as a sample from the joint posterior distribution. Due to the orthogonal invariance of the posterior distribution, which results from the orthogonal invariance of the chosen prior distribution and the orthogonally invariant likelihood, the posterior distribution may undergo an orthogonal transformation as in Equation 4.8, see also Chapter 3.

The second step of the proposed approach therefore consists of finding an orthogonal matrix D to transform the posterior distribution by. The matrix D is chosen in such a way that the multivariate highest posterior density intervals (HPDIs) or highest posterior density ellipsoids (HPDEs) for Λ of appropriately chosen width contain the zero either for a particular subset of elements of Λ, or simply for as many elements of Λ as possible. The former approach can be used in order to find an interpretable structure of Λ, and the latter approach can be used to find the most parsimonious structure. Eventually, a CFA is conducted conditional on the identified parsimonious structure. In this sense, the approach resembles the model selection

[11] In this light, the sequences from the scenario with 55% sparsity in Λ may also be interpreted as switching between different modes, rather than slowly converging.

procedure proposed by Millsap (2001) described in Section 4.2.3, where the result of an EFA is used as the basis for a subsequent CFA. The main difference here lies in the use of HPDEs to determine the associations between the variables and the factors instead of using point estimates for Λ only.

4.4.1 The Weighted Orthogonal Procrustes Step

In order to obtain a sample from the posterior distribution of Λ, from which the multivariate HPDIs can be constructed, the unconstrained Gibbs sampler as described in Appendix 4.A is run. Appendix 4.A also describes the chosen prior distributions for the parameters and the requirements they have to satisfy for the unconstrained Gibbs sampler. The Gibbs sampler, see e.g. Casella and George (1992), iteratively draws from the full conditional distributions of the loadings Λ and of the idiosyncratic covariance matrix Σ, which is diagonal. The factors $\{f_t\}_{t=1}^T$ are augmented parameters and are accordingly drawn from their full conditional distribution, see Tanner and Wong (1987) for the general concept of data augmentation, and Otrok and Whiteman (1998) for augmenting factors in Bayesian factor analysis. The sampler's output with respect to each set of parameters converges to samples from the marginal distributions of these parameters, see e.g. Geman and Geman (1984) or Gelfand and Smith (1990). Note that no PLT constraint is imposed on Λ, hence the rotation problem is unsolved and the obtained sequences of Λ and the factors are orthogonal mixtures, due to the unconstrained sampler's *orthogonal mixing*.[12] In an orthogonally mixing sampler, at least some parameters of the full conditional distributions undergo orthogonal transformations during the sampling process. The effect of such an orthogonal transformation on the parameters of the full conditional distribution is explained in Appendix 4.B.

Due to the fact that the product of two orthogonal matrices is itself an orthogonal matrix, the successive transformations of the sample space of Λ in an arbitrarily long sequence of draws can be expressed as a single orthogonal matrix. If a sequence of Z draws from the unconstrained Gibbs sampler is at hand, reversing the orthogonal mixing therefore requires that for each draw, an orthogonal matrix $D^{(z)}$ is found that Λ and $\{f_t\}_{t=1}^T$ are transformed by to remove the orthogonal mixing and ensure that after the transformation, the sample space is the same for all elements of the sample. The $D^{(z)}$ are found by means of the WOP algorithm described in Appendix 4.C and are used to map all the $\Lambda^{(z)}$ for $z \in \{1, \ldots, Z\}$ into the same sample space for Λ. This sample space, however, may be replaced by an arbitrary orthogonal transformation of the same sample space.[13] Accordingly, the entire postprocessed sequences of loadings and factors may undergo this orthogonal transformation. Such a transformation can be chosen to facilitate the interpretation of the postprocessed Gibbs output. For instance, a rotation technique like Varimax, see Kaiser (1958), Quartimax, see Neuhaus and Wrigley

[12] Orthogonal mixing can be seen as a continuous version of label switching, see e.g. Stephens (2000) or Jasra et al. (2005).

[13] In fact, the orientation of the sample space of Λ depends on the initialization of the WOP approach, where, for simplicity, the last draw from the sampler $\Lambda^{(Z)}$ is used.

(1954), or an approach designed to find a specific pattern in the loadings matrix, as in Boysen-Hogrefe and Pape (2011), can be used.

Assuming that the orthogonal mixing has been successfully removed by the WOP approach, the marginal posterior distribution of Λ is elliptical for every row of Λ, i.e. λ_i for $i \in \{1, \ldots, N\}$, as the likelihood is elliptical, being a normal distribution, and the prior for Λ is also elliptical, being a product of normal distributions. This property is important in the next step of the proposed approach to find a parsimonious loadings structure.

4.4.2 The Sparse Pattern Identification Step

Having obtained a sample from the marginal posterior distribution of Λ postprocessed with the WOP algorithm and denoted $\{\Lambda^{(z)}\}_{z=1}^{Z}$, I next determine which of the elements of the loadings matrix Λ can be assumed to be zero. The Bayesian approach corresponding to hypothesis testing is to look at the $1-\alpha$ highest posterior density intervals (HPDIs) of the posterior distribution of the parameters of interest, where α is the level of the "test", see e.g. Hoff (2009). This is fairly easy for univariate posterior distributions, particularly if they are unimodal and symmetric, but can be troublesome if multiple parameters are of interest and the posterior distribution is therefore multivariate. Fortunately, the posterior distribution of Λ has two invaluable properties: First, it is elliptical, and second, it can be orthogonally transformed as required.

Scheffé (1953) discusses the issue of multivariate confidence intervals, or confidence spheres, and Hanson and McMillan (2012) suggest similar approaches to obtain multivariate HPDIs in Bayesian analysis. The third approach suggested in Hanson and McMillan (2012) follows the construction of multivariate confidence intervals in Hauck (1983), which is suitable for elliptical distributions and works with the Mahalanobis distance. It is also suitable here due to the ellipticity of the marginal posterior distributions of λ_i for $i \in \{1, \ldots, N\}$ once the orthogonal mixing has been removed. The multivariate HPDIs obtained from elliptical posterior distributions are consequently also elliptical, and can therefore also be considered as highest posterior density ellipsoids (HPDEs).

The estimated Mahalanobis distance of each $\lambda_i^{(z)}$ from the (estimated) center of its distribution is

$$d_i^{(z)} = (\lambda_i^{(z)} - \widehat{E(\lambda_i)})' \hat{\Sigma}_i^{-1} (\lambda_i^{(z)} - \widehat{E(\lambda_i)}) \quad \text{for} \quad z \in \{1, \ldots, Z\}, \tag{4.36}$$

where $\widehat{E(\lambda_i)} = Z^{-1} \sum_{z=1}^{Z} \lambda_i^{(z)}$ and $\hat{\Sigma}_i = \widehat{\mathrm{Cov}(\lambda_i)} = Z^{-1} \sum_{z=1}^{Z} \lambda_i^{(z)} \lambda_i^{(z)\prime} - \widehat{E(\lambda_i)}\widehat{E(\lambda_i)}'$.

To obtain the $1-\alpha$ HPDEs for each λ_i, the $\lfloor \alpha Z \rfloor$ points located at the greatest Mahalanobis distance from the mean are discarded for each λ_i. The remaining sample is then $\{\Lambda^{(\tilde{z})}\}_{\tilde{z}=1}^{\tilde{Z}}$, where $\tilde{Z} = \lceil (1-\alpha)Z \rceil$. Since the Mahalanobis distance in Equation (4.36) is invariant to a joint orthogonal transformation of $\{\Lambda^{(z)}\}_{z=1}^{Z}$, such a transformation simply results in a change in the orientation of all N HPDEs, which otherwise retain their shape. Hence the

same HPDEs are obtained if the steps of orthogonally transforming $\{\Lambda^{(z)}\}_{z=1}^{Z}$ by a matrix D and discarding the $\lfloor \alpha Z \rfloor$ outermost points are processed in reverse order. Thus it is possible to choose an arbitrary orientation of the sample, obtain the HPDEs and apply a joint orthogonal transformation to the HPDEs afterwards. This transformation can be chosen such that e.g. as many zeros as possible are included in the HPDEs, which indicates a parsimonious structure in Λ.

Any orthogonal transformation that is jointly applied to all N HPDEs can be expressed as a combination of three types of orthogonal transformations: rotations, axis reflections and permutations. Rotations can be further decomposed into Givens rotations, axis reflections into single-axis reflections, and permutations into pairwise permutations, which are described in the following.[14]

A Givens rotation is a rotation around two axes. Any rotation around K axes with $K \geq 2$ can be expressed as a combination of $\binom{K}{2}$ Givens rotations, or as the product of the respective Givens rotation matrices, see e.g. Bernstein (2009). Since Givens rotations always involve two axes, the remaining axes stay unaffected. It is therefore possible to proceed sequentially and consider one pair of axes at a time, which corresponds to one pair of columns in Λ. The zeros found in the HPDEs for the columns of Λ not involved in the Givens rotation then stay the same. Consider e.g. a rotation around the axes k and l: If the i^{th} HPDE does not include the zero for $\lambda_{i,k}$ initially, the rotated HPDE may do so for $\lambda_{i,k}$. For some $\lambda_{j,l}$, however, the adverse effect may be observed: The j^{th} HPDE may include the zero, but this may not be the case for the rotated HPDE. In other words, reducing one element of the loadings matrix to zero may result in other elements to be different from zero afterwards. A reflection about one axis changes the signs of the corresponding column vector of Λ. Reflections about multiple axes are then simply a combination of up to K axis reflections, or the product of the respective axis reflection matrices. Note, however, that if the i^{th} HPDE includes the zero for $\lambda_{i,k}$, it still does so after the reflection about the k^{th} axis, so this type of orthogonal transformation can be neglected. A pairwise permutation exchanges a pair of columns of Λ. Any permutation of $K \geq 2$ columns can be expressed as a combination of $\binom{K}{2}$ pairwise permutations, or as the product of the respective permutation matrices. If the i^{th} HPDE includes the zero for $\lambda_{i,k}$, however, after a permutation of factor k and l, it contains the zero for $\lambda_{i,l}$ instead, so this type of transformation can also be neglected. It may be useful to transform the results by means of reflections and permutations for purposes of interpretability, however, the set of zero elements in the loadings matrix is not affected by these types of orthogonal transformations. Therefore I focus exclusively on Givens rotations about all pairs of axes.

[14] In fact, as Theorem 2.4.1 shows, any K-dimensional orthogonal transformation can be expressed as a combination of at most $\binom{K}{2}$ Givens rotations and a reflection about the K^{th} axis. As it is argued in the following that the only relevant orthogonal transformations for the proposed approach are rotations, the theorem is not necessary here.

Regarding the question of multimodality discussed in Section 4.3.4, applying rotations to the HPDEs allows to explore the possible multiple sparse representations of Λ. If an orthogonal transformation by a matrix D_1 is found that maximizes the number of zero elements in Λ, there may exist a second orthogonal transformation by a matrix $D_2 \neq D_1$, which results in the same number of zeros in Λ, albeit in different places. Whereas the sparse sampler in Section 4.3.2 can be observed to "lock in" to one sparse representation of Λ, the proposed approach allows to explore the possible multiple sparse representations.

Next, I define an auxiliary function that measures the location of the HPDEs. It determines for each K-variate HPDE and each of the K axes involved in the HPDE the share of points with positive and negative values for $\lambda_{i,k}$, respectively. Geometrically speaking, it measures for each axis the share of points that lies on each side of zero. For instance, for some $\lambda_{i,k}$ 80% of the points lie below zero and 20% lie above zero. Only the smaller of these two figures is considered, which is strictly positive if the zero is included in the HPDE for $\lambda_{i,k}$. Geometrically speaking, there must be at least one point on the axis located on the positive and one point on the negative side of zero on the k^{th} axis of the i^{th} HPDE to ensure that the zero for $\lambda_{i,k}$ is also included in the HPDE.

The auxiliary function is

$$\kappa_{i,k} = \kappa_{i,k}(\{\Lambda^{(\tilde{z})}\}_{\tilde{z}=1}^{\tilde{Z}}, D) = \min\left(\sum_{\tilde{z}=1}^{\tilde{Z}} \frac{I_{\{(\Lambda^{(\tilde{z})}D)_{i,k}>0\}}}{\tilde{Z}}, \sum_{\tilde{z}=1}^{\tilde{Z}} \frac{I_{\{(\Lambda^{(\tilde{z})}D)_{i,k}<0\}}}{\tilde{Z}}\right), \tag{4.37}$$

where $(\Lambda^{(\tilde{z})}D)_{i,k}$ denotes the element in the i^{th} row and the k^{th} column of the $\Lambda^{(\tilde{z})}$ which has been transformed by the orthogonal matrix D. In Equation (4.37), it must therefore hold that $\kappa_{i,k}(\cdot) > 0$ in order to set $\lambda_{i,k}$ to zero. This, however, is a necessary, but not a sufficient condition. Consider a case where the HPDE contains the zero either for one axis or for a different axis, but not for both at the same time. This case is shown for a bivariate HPDE in Figure 4.12. Thus when the set of elements of λ_i for which the HPDE contains the zero has been found, the next task is to find the largest subset of this set for which the HPDE contains the zero at the same time. Generally, let $\mathcal{K} \subseteq \{1, \ldots, K\}$ denote a subset of the elements 1 to K. Then the auxiliary function can be generalized for this subset as

$$\kappa_{i,\mathcal{K}} = \kappa_{i,\mathcal{K}}(\{\Lambda^{(\tilde{z})}\}_{\tilde{z}=1}^{\tilde{Z}}, D) = \min\left(\sum_{\tilde{z}=1}^{\tilde{Z}} \frac{I_{\{(\Lambda^{(\tilde{z})}D)_{i,\mathcal{K}}>\mathbf{0}\}}}{\tilde{Z}}, \sum_{\tilde{z}=1}^{\tilde{Z}} \frac{I_{\{(\Lambda^{(\tilde{z})}D)_{i,\mathcal{K}}<\mathbf{0}\}}}{\tilde{Z}}\right), \tag{4.38}$$

where $\mathbf{0}$ is a vector of zeros of the same length as $\mathcal{K}$.

The task of finding the largest subset $\mathcal{K}$ for which $\kappa_{i,\mathcal{K}} > 0$ becomes much easier by the pairwise processing of the axes, or the columns of Λ. This is possible because the K-dimensional rotation matrix D can be replaced by up to $\binom{K}{2}$ Givens rotation matrices.

The following algorithm describes the approach to find the most parsimonious structure in Λ, whose pattern is collected in the $N \times K$ indicator matrix Δ.[15] At initialization, all elements of Δ are set to 1 and $D = I_K$. The pair of axes that the Givens rotations in Step 4 are applied to is denoted as $\mathfrak{K} = \{\mathfrak{k}_1 \;\; \mathfrak{k}_2\} \subseteq \{1, \ldots, K\}$, where initially, $\mathfrak{K} = \{1, 2\}$.

Algorithm 4.4.1.

1. *Given D and $\mathfrak{K}$, calculate $\kappa_{i,\mathfrak{k}_1}$ and $\kappa_{i,\mathfrak{k}_2}$ as in Equation (4.37) for every $i \in \{1, \ldots, N\}$.*
2. *Obtain for every $i \in \{1, \ldots, N\}$ the subset of $\mathfrak{K}$ for which the condition $\kappa_{i,\mathfrak{k}} > 0$ holds, denoted as the set $\mathcal{K}_i \subseteq \mathfrak{K}$.*
3. *For every $i \in \{1, \ldots, N\}$ where $\mathcal{K}_i \neq \emptyset$, determine which of the following four cases applies:*
   - 3a. *For every $i \in \{1, \ldots, N\}$ where $\mathcal{K}_i = \mathfrak{K}$, i.e. the zero is simultaneously included in the HPDE for both considered columns $\mathfrak{k}_1$ and $\mathfrak{k}_2$, set $\delta_{i,\mathfrak{k}_1}$ and $\delta_{i,\mathfrak{k}_2}$ to zero.*
   - 3b. *For every $i \in \{1, \ldots, N\}$ where $\mathcal{K}_i = \mathfrak{k}_1$, i.e. the zero is included in the HPDE only for column $\mathfrak{k}_1$, set $\delta_{i,\mathfrak{k}_1} = 0$.*
   - 3c. *For every $i \in \{1, \ldots, N\}$ where $\mathcal{K}_i = \mathfrak{k}_2$, i.e. the zero is included in the HPDE only for column $\mathfrak{k}_2$, set $\delta_{i,\mathfrak{k}_2} = 0$.*
   - 3d. *For every $i \in \{1, \ldots, N\}$ where $\mathcal{K}_i \neq \mathfrak{K}$, but $\kappa_{i,\mathfrak{k}_1} > 0$ and $\kappa_{i,\mathfrak{k}_2} > 0$, which implies that the zero is included in the HPDE either for column $\mathfrak{k}_1$ or for column $\mathfrak{k}_2$, but not for both at the same time, set either $\delta_{i,\mathfrak{k}_1}$ or $\delta_{i,\mathfrak{k}_2}$ to zero, where each of the two is chosen with probability $\frac{1}{2}$.*
4. *Calculate $\eta = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{i,k}$, where only the elements in the columns $\mathfrak{k}_1$ and $\mathfrak{k}_2$ can have changed since the last iteration.*
5. *While η can be decreased, apply a different Givens rotation around axes $\mathfrak{k}_1$ and $\mathfrak{k}_2$ to D and restart from Step 1.*
6. *Change the pair of axes $\mathfrak{K} = \{\mathfrak{k}_1, \mathfrak{k}_2\}$ while any two-element subset of $\{1, \ldots, K\}$ has not been considered yet, and proceed with Step 1.*

Regarding Algorithm 4.4.1, it must be noted that if $\kappa_{i,\mathfrak{K}} > 0$, i.e. zero is included in the HPDE for $\lambda_{i,\mathfrak{k}_1}$ and $\lambda_{i,\mathfrak{k}_2}$ at the same time, a rotation around the axes $\mathfrak{k}_1$ and $\mathfrak{k}_2$ has no effect on $\kappa_{i,\mathfrak{K}}$, as the point $(\mathfrak{k}_1, \;\; \mathfrak{k}_2)' = (0, \;\; 0)'$ necessarily remains within the HPDE. A rotation about a different subset of $\{1, \ldots, K\}$, however, may change this.

$\eta = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{i,k}$ denotes the number of ones in Δ, and consequently, the number of nonzero elements in Λ. Denote $\Delta(D)$ as the indicator matrix Δ that results from the orthogonal

[15] The elements of Δ, which take only binary values, can be interpreted as association probabilities in the subsequently conducted confirmatory factor analysis.

transformation matrix D obtained from Algorithm 4.4.1. Then, the most parsimonious structure in Λ is identified by

$$D^* = \arg\min_D \sum_{i=1}^{N} \sum_{k=1}^{K} (\Delta(D))_{i,k}. \tag{4.39}$$

If instead of the most parsimonious representation, a particular pattern in Λ is of interest, the $(\Delta(D))_{i,k}$ can multiplied by positive or negative weights to reward or punish zero elements in particular places. Two examples for this are given below, where the proposed algorithm is applied to answer the two questions asked at the beginning of this chapter, regarding the number of factors and the choice of relevant variables.

The optimal D, denoted as D^*, can be found by means of a nonlinear global optimization algorithm, such as the Nelder-Mead algorithm, see Nelder and Mead (1965). The optimization routine can be sped up the by adding a continuous term to the term in Equation (4.39), obtaining the augmented function

$$D^* = \arg\min_D \sum_{i=1}^{N} \sum_{k=1}^{K} \left((\Delta(D))_{i,k} \left\{ 1 + (0.5 - \kappa_{i,k}) \right\} \right). \tag{4.40}$$

The choice of the additive term $0.5 - \kappa_{i,k}$ is based on the fact that $\kappa_{i,k}$ can reach at most a value of 0.5, which is the case if the i^{th} HPDE is exactly centered about zero for the k^{th} axis. Thus, the term increases as mass is shifted to either side, making the i^{th} HPDE "just" contain zero for the k^{th} axis. Loosely speaking, this allows the HPDE to reach out further in other directions, possibly including zeros on additional axes.[16]

The three questions mentioned in Section 4.1, dealing with the identification of irrelevant variables, determining the number of factors and identifying a sparse loadings pattern, can now be tackled in different ways by applying Algorithm 4.4.1. The objective function in Equation (4.40) minimizes the total number of nonzero elements in the loadings matrix. Consider the question whether the number of factors is correctly specified or chosen too large. The objective function can thus be changed to

$$D^* = \arg\min_D \sum_{i=1}^{N} \sum_{k=1}^{K} \left((\Delta(D))_{i,k} \left\{ 1 + (0.5 - \kappa_{i,k}) + I_{\left\{ \left(\sum_{j=1}^{N} (\Delta(D))_{j,k} \right) < 3 \right\}} \right\} \right), \tag{4.41}$$

i.e. zero elements that are located in a column with less than three nonzero elements get twice the weight of the other zero elements. As the objective is to find many zero elements in the same column, Step 3d. of Algorithm 4.4.1 should be accordingly adjusted, ensuring that whenever two possible choices for $\delta_{i,k}$ to be set to zero exist with respect to k, the column already containing more zero elements is chosen. Note that doubling the weights is a heuristic decision; other choices may likewise be justified. It turns out in the simulation study, however,

[16] It must be noted, however, that this increases the risk of incurring cases where zero is contained in the HPDE for multiple axes individually, but not simultaneously. Thus an increase in $0.5 - \kappa_{i,k}$ by a small amount may be accompanied by a decrease in $(\Delta(D))_{i,k}$ by one, due to the loss of a zero element.

that the double weight allows to identify a misspecification in the number of factors - i.e. cases where the number of factors has been chosen too large - very well.

The elimination of a spurious factor in the case of less than three nonzero elements is based on the three-indicator rule for confirmatory factor analysis with uncorrelated factors, see e.g. Bollen (1989), p.274. This rule builds on Theorem 5.5 from Anderson and Rubin (1956) that states as a sufficient condition to identify a factor, three loadings must be different from zero.[17] In this case, the optimization therefore focuses on setting an entire column of loadings, except for two or fewer elements, to zero.

Regarding the identification of irrelevant variables, no optimization is necessary at all: If $\sum_{k=1}^{K}(\Delta(D))_{i,k} = 0$ holds, this implies that the origin for λ_i is included in the i^{th} HPDE. This cannot be changed by a rotation, see also Kaufmann and Schumacher (2012), so no alternative choices for D have to be considered. To find out whether a variable can be eliminated from the sample, Algorithm 4.4.1 must therefore only evaluate Δ for the initial $D = I_K$ and determine whether any row of the resulting indicator matrix Δ contains only zeros.

Having found the desired parsimonious representation of Λ, which is then stored in the indicator matrix Δ, a Bayesian confirmatory factor analysis is run, where the loadings corresponding to zero elements in Δ are set to zero. The Bayesian confirmatory factor analysis uses a Gibbs sampler similar to the one in Appendix 4.A, but with the prior for Λ accordingly adjusted, such that a Dirac Delta prior is used for $\lambda_{i,k}$ where $\delta_{i,k} = 0$. Consequently, only the nonzero elements of every λ_i are sampled. Variables with exclusively zero loadings, corresponding to entire rows of zeros in Δ, can be omitted altogether. In general, no additional identification constraints are necessary in the confirmatory factor analysis. If the Gibbs sampler output displays orthogonal mixing in the form of reflections, the output can be postprocessed e.g. by the algorithm proposed by Erosheva and Curtis (2013) for Bayesian confirmatory factor analysis, compare also Example 2.5.6 in Chapter 2. The confirmatory factor analysis is performed after identifying the sparse pattern both in the simulation study in Section 4.5 and in the empirical application in Section 4.6.

4.5 Three Experiments

To evaluate the capabilities of the proposed two-step approach, I generate artificial data following the simulation setup of Lopes and West (2004) and Frühwirth-Schnatter and Lopes (2012). In all three experiments, 50 data sets are simulated. The setup for all three experiments is similar, with a sparse $N \times K$ loadings matrix and orthogonal static factors with unit variance each, where $T = 100$.

[17] For correlated factors, an according two-indicator rule exists.

4.5.1 Setups, Estimation Procedure and Benchmarks

Experiment 1 is the baseline scenario, where the loadings matrix used in the simulation of the data has a congeneric structure and is

$$\Lambda' = \begin{pmatrix} 0.99 & 0 & 0 & 0.99 & 0.99 & 0 & 0 & 0 & 0 \\ 0 & 0.95 & 0 & 0 & 0 & 0.95 & 0.95 & 0 & 0 \\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0.9 & 0.9 \end{pmatrix}. \tag{4.42}$$

The corresponding idiosyncratic covariance matrix is

$$\Sigma = \text{diag}\begin{pmatrix} 0.02 & 0.19 & 0.36 & 0.02 & 0.02 & 0.19 & 0.19 & 0.36 & 0.36 \end{pmatrix}. \tag{4.43}$$

In this experiment, K is correctly specified in the analysis.

Experiment 2, which focuses on the identification of irrelevant variables, resembles the baseline scenario, but contains three additional observable variables, which have zero loadings on all the factors, so

$$\Lambda' = \begin{pmatrix} 0.99 & 0 & 0 & 0.99 & 0.99 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.95 & 0 & 0 & 0 & 0.95 & 0.95 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0.9 & 0.9 & 0 & 0 & 0 \end{pmatrix} \tag{4.44}$$

and

$$\Sigma = \text{diag}\begin{pmatrix} 0.02 & 0.19 & 0.36 & 0.02 & 0.02 & 0.19 & 0.19 & 0.36 & 0.36 & 1 & 1 & 1 \end{pmatrix}. \tag{4.45}$$

This implies that the three additional variables contain only noise, but no systematic information. Following the argumentation of Boivin and Ng (2006) for exploratory factor models and Kaufmann and Schumacher (2012, 2013) for sparse factor models, it is useful to identify and remove such irrelevant variables. As in the first experiment, K is correctly specified here.

Experiment 3 uses the data from the baseline scenario again. The sampler, however, is run with a misspecified K here, so K is determined endogenously by starting from some value $\hat{K}_{max} > K$, and reducing the number of factors until no further reduction is possible. As long as it is possible to discard a spurious factor by means of this procedure, the number of factors is still specified too large. Hence it makes sense to start with K_{max} larger than the expected K, since conclusions about K can be drawn only if factors can be eliminated from the initial estimation setup. If none can be eliminated, the number of factors has either been chosen correctly - or too small, if the model has not been estimated with one additional factor previously.

For all three experiments, the following procedure is applied: Prior to the Bayesian analysis, a PC factor analysis is performed, and the estimated loadings matrix obtained therefrom is then rotated according to the Varimax criterion by Kaiser (1958). The resulting estimate may

provide a first insight into the structure of the loadings matrix, compare e.g. Eickmeier (2005) or the algorithm by Kaufmann and Schumacher (2013) described in Section 4.3.2, where the Varimax solution from a PC factor analysis serves as the initialization of the sampler.

Next, the unconstrained Gibbs sampler is run, choosing as prior hyperparameters $\underline{\alpha}_i = 1$, $\underline{\beta}_i = 1$ and $\underline{c}_i = 1$ for $i \in \{1, \ldots, N\}$, for an initial burn-in phase of 2,000 iterations. Afterwards, the sampler is run for another 10,000 iterations. The statistic by Geweke (1992) is used to monitor convergence of the sampler, which resembles the test statistic in a mean difference test. For a quantity of interest x, where a realized sample $\{x^{(z)}\}_{z=1}^{Z}$ is observed, the convergence statistic is

$$Q = \frac{(\overline{x}_1 - \overline{x}_2)^2}{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}, \tag{4.46}$$

where $\overline{x}_1$ is the mean of one subsection of the sample of length n_1 and $\overline{x}_2$ is the mean of one subsection of the sample of length n_2. $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the corresponding variance estimates. Asymptotically, Q follows a χ^2-distribution with 1 degree of freedom. The convergence statistic checks whether the mean at the beginning and at the end of the observed sequence are significantly different. Thus, $\overline{x}_1 = \frac{1}{\lfloor 0.2Z \rfloor} \sum_{z=1}^{\lfloor 0.2Z \rfloor} x^{(z)}$ is the mean over the first 20% of the sample, and $\overline{x}_2 = \frac{1}{\lfloor 0.5Z \rfloor} \sum_{z=\lfloor 0.5Z \rfloor + 1}^{Z} x^{(z)}$ is the mean over the last 50% of the sample.

Autocorrelation in the Gibbs sequence can be accounted for by using autocorrelation-robust variance estimates in Equation (4.46). A covariance estimator robust against heteroscedasticity and autocorrelation has been proposed by Newey and West (1987) is

$$\hat{S} = \hat{\Gamma}(0) + \sum_{\tau=1}^{l} \left(1 - \frac{\tau}{l+1}\right) (\hat{\Gamma}(\tau) + \hat{\Gamma}(\tau)'), \tag{4.47}$$

where $\hat{\Gamma}(\tau)$ denotes the estimated autocovariance matrix of the Gibbs sequence of order τ, see Hamilton (1994), Chapter 10.5. The number of lags l can either be determined by looking at the autocorrelation function, or, as rule of thumb, be chosen as $\lceil \tilde{n}^{\frac{1}{4}} \rceil$, where $\tilde{n}$ is the length of the partial sequence, i.e. either n_1 or n_2. A univariate version of Equation (4.47) then yields $\hat{s}_1^2$ and $\hat{s}_2^2$, which replace $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ in Equation (4.46).

Since the sampler is orthogonally mixing, convergence of orthogonally invariant quantities must be monitored. The idiosyncratic variances are orthogonally invariant, but instead of the $\lambda_{i,k}$, the sum of squared loadings per row vector of Λ are monitored, i.e. $\sum_{k=1}^{K} \lambda_{i,k}^2$. Convergence is assumed if the tests indicate convergence at the 5% level for at least 90% of the monitored quantities. If no convergence is attained, the length of the burn-in sequence is extended by 1,000, and the retained sequence keeps its initial length.

The converged sample from the directionally unconstrained sampler is then postprocessed with the WOP approach to obtain the posterior densities identified up to the choice of D, from which the initial HPDEs for every λ_i with $i \in \{1, \ldots, N\}$ are constructed. To identify a parsimonious structure in Λ for Experiment 1, D is chosen as the D^* obtained from Equation (4.40). In Experiment 2, the initial HPDEs already indicate which variables are irrelevant - an orthogonal transformation of the HPDEs does not change the zero rows in Δ. Nonetheless, D^* is obtained from Equation (4.40) in order to find a parsimonious loadings structure for the relevant variables as well. In Experiment 3, D^* is obtained from Equation (4.41). The initial HPDEs are constructed as described in Section 4.4, using three different values for α, namely $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. In all three experiments, a confirmatory Bayesian factor analysis is run based on the identified sparse structure in Λ with the same prior hyperparameter settings and the same convergence checks to obtain a sequence of 10,000 draws from the posterior distribution. The results are compared to the outcome of the initial PC factor analysis with factor loadings transformed to maximize the Varimax criterion, to the outcome of the WOP approach, and to the outcome of a sparse factor analysis using the sampler described in Section 4.3.2.

4.5.2 Results for the Two-Step Approach and the Benchmarks

For **Experiment 1**, the degree of sparsity is determined by the proposed two-step approach for three different choices of α in constructing the HPDEs and by the sparse sampler from Section 4.3.2 with the prior hyperparameter settings as in Section 4.3.4, i.e. $\underline{\alpha}_i = 1$ and $\underline{\beta}_i = 1$ for all $i \in \{1, \ldots, N\}$, $c_0 = 2$ and $C_0 = 0.5$, and $s_k = 1$, $r_k = 0.8$, $w_k = 500$ and $v_k = 0.1$ for all $k \in \{1, \ldots, K\}$. The results for the 50 simulations are reported in Table 4.3 in terms of relative frequencies of occurrence for each number of nonzero elements. It appears reasonable that, using wider intervals, i.e. smaller α values, allows to include more elements of Λ in the HPDEs, resulting in a higher share of successfully identified models. In the toy example used in the simulation study, there is no caveat to the use of very small α values, since the elements of Λ in Equation (4.42) are either zero or large in magnitude. For empirical data sets, this may not be the case, and the results may critically depend on the choice of α.

Table 4.4 reports the mean estimates for Λ from the confirmatory factor analysis that is based on the sparse structure identified by the two-step approach. The results have been adjusted for column permutations and reflections to take the structure of Λ that was used in simulating the data, given in Equation (4.42). As α increases, chances are higher that zero elements in the loadings matrix are not identified. As the results are similar for all three choices of α, the statistical accuracy of the estimates for Λ does not seem to depend on α. For comparison, Table 4.5 reports the results of a PC factor analysis with a subsequent Varimax rotation, the WOP results, and the estimate for Λ from sparse sampler in Section 4.3.2. The sparse sampler reaches a similar degree of sparsity as the two-step approach does for $\alpha = 0.05$. The estimates for the nonzero elements of Λ, however, are quite different from each other, and also differ from the results for the two-step approach.

For **Experiment 2**, Table 4.6 shows the relative frequencies for each number of rows with nonzero elements in the upper part, and the relative frequencies for each number of total nonzero elements in the lower part. The two-step approach identifies the irrelevant variables in all cases for $\alpha = 0.01$ and $\alpha = 0.05$, and in almost all cases for $\alpha = 0.1$. The sparse sampler fares slightly worse. As can be seen in the lower section of the table, the two-step approach with $\alpha = 0.1$ tends not to find all zero elements in Λ anymore, which may be due to the irrelevant variables and could thus be seen as an argument for not including them in the analysis.[18] Table 4.7 shows that for $\alpha = 0.01$ and $\alpha = 0.05$, the factor loadings patterns for the first nine variables are almost identical to the ones from the first experiment, and the patterns for the last three variables show that these variables are overall correctly identified as not loading on any of the factors. For $\alpha = 0.1$, some zero loadings in Λ are occasionally not correctly identified. Table 4.8 reports the results analogously for a PC factor analysis with a subsequent Varimax rotation, the WOP results, which have been orthogonally transformed to minimize the distance to the matrix in Equation (4.42), and the estimate for Λ from sparse sampler from Section 4.3.2. The estimates are similar to those in Table 4.5, however, the sparse sampler performs slightly worse than the two-step approach in identifying the irrelevant variables.

For **Experiment 3**, Table 4.9 reports the number of factors found by transforming the HPDEs by the matrix D^* determined according to the criterion in Equation (4.41) in the upper part, and the relative frequencies for the number of nonzero elements in the lower part. The two-step approach, starting from $K_{max} = 4$ and reducing the number of factors, and the sparse sampler from Section 4.3.2 correctly determine $K = 3$ in all cases. The relative frequencies for the number of nonzero elements that is found resembles that in Experiment 1. Table 4.10 shows the mean estimates for Λ from the confirmatory factor analysis, which are similar to those from Experiment 1. The results for the PC factor analysis with a subsequent Varimax rotation, the WOP results, and the estimate for Λ from sparse sampler in Section 4.3.2 shown in Table 4.11 are also similar to those from the first experiment throughout.

Table 4.12 shows the root mean-squared errors (RMSEs) for the nonzero elements of Λ and the diagonal elements of Σ in the three experiments, calculated for each of the nine relevant elements of Λ and Σ as

$$\xi(\lambda) = \sqrt{\frac{1}{50}\sum_{i=1}^{50}(\hat{\lambda} - \lambda)^2} \qquad \text{and} \qquad \xi(\sigma^2) = \sqrt{\frac{1}{50}\sum_{i=1}^{50}(\hat{\sigma}^2 - \sigma^2)^2}. \tag{4.48}$$

Instead of reporting the RMSE for all nine elements, the RMSEs are arranged in increasing order of magnitude and every other of them is reported, starting with the first. Interestingly, if the number of factors is correctly specified, i.e. in Experiment 1 and 2, the PC factor analysis with subsequent Varimax rotation and the WOP approach yield virtually identical RMSEs. In Experiment 3, where the number of factors is not correctly specified, the results from WOP differ from those of the PC factor analysis with subsequent Varimax rotation. Regarding the

[18] The change is small, however, so the finding may be purely coincidental.

RMSEs of the σ^2, the sparse sampler and the two-step approach yield almost identical results, whereas the two-step approach achieves much smaller RMSEs than the sparse sampler for the elements of Λ. Both samplers, however, fare worse than WOP and the PC factor analysis with subsequent Varimax rotation. Note, however, that the two latter approaches require a full Λ matrix, whereas in the former two approaches, up to 75% of the elements of Λ are set to zero.

The three experiments show that for the simulated data, the two-step approach performs very well, both with respect to answering the three initial questions, i.e. determinining the number of factors, identifying irrelevant variables and finding a parsimonious structure in the loadings matrix, and with respect to estimating the remaining model parameters. Hence the approach could provide helpful insights in empirical applications. The following section will check this, using a data set that has often been analyzed with different types of factor models.

4.6 Empirical Application: Students Test Data

The data set used in the empirical application is from the bi-factor study by Holzinger and Swineford (1939). It consists of the observed test scores for 301 students from two high schools, who were given the 24 different tasks listed in Table 4.13.

The bi-factor model assumes a structure where each variable loads on a general and a group factor (Holzinger and Swineford, 1937).[19] The general factor in this model is understood as a "general intelligence" factor, beside which Holzinger and Swineford (1939) assume the existence of five group factors. Item 1 to 4 should load on the first group factor, the "spatial" factor, item 5 to 9 on the second group factor, the "verbal" factor, item 10 to 13 on the third group factor, the "speed" factor, item 14 to 19 on the fourth group factor, the "memory" factor, and item 20 to 24 on the fifth group factor, the "mathematical deduction" factor. In their original study, Holzinger and Swineford (1939) cannot establish the fifth group factor as orthogonal to the remaining ones and therefore drop it.

The purpose of this application is to analyze the data set without assuming any knowledge about the number of factors, whether all variables are relevant and how the parsimonious, or sparse, loadings structure looks like. Such a task can be performed based on PC factor analysis with a subsequent rotation to the bi-factor structure. It is also possible to run a Bayesian factor analysis with the WOP approach and rotate the solution as close as possible to the supposed bi-factor structure. Eventually, the sparse sampler from Section 4.3.2 can be applied to find a parsimonious structure in Λ. As in the simulation study, these three approaches are used as benchmarks for the two-step procedure. The two-step procedure is performed for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Starting with a sufficiently large number of factors $K_{max} = 6$, I first try to reduce the number of factors by using the optimization criterion in Equation (4.41).

[19] The bi-factor setup first appears in psychometric context in the work of Holzinger and Swineford (1937, 1939), but more recently, a dynamic factor model with a bi-factor loadings structure for macroeconomic data has been estimated by Kose et al. (2003), who accordingly determine a world and several regional business cycle factors and use Bayesian factor analysis as the estimation approach.

After completing this step, I re-estimate the model with the correctly specified K and identify irrelevant variables, which do not load on any of the factors. Eventually, I use only the relevant variables, estimate a model with the correctly specified number of factors and try to establish a parsimonious structure with the criterion in Equation (4.40). As the difference between zero and nonzero loadings is probably not as clear-cut as for the simulated data, the different choices of α may yield different parsimonious structures, and possibly also different numbers of factors. Sufficiently "strict" criteria, i.e. very small values of α may remove variables that are considered relevant if higher values for α are chosen. To analyze the robustness of the results obtained from the two-step approach, the analysis is repeated 20 times, where the Gibbs sampler uses a different seed each time.[20] The prior hyperparameters are again $\underline{\alpha}_i = 1$, $\underline{\beta}_i = 1$ and $\underline{c}_i = 1$ for $i \in \{1, \ldots, N\}$, and convergence is again monitored as before, except that a sequence of length 20,000 from the posterior distribution is generated. The sparse sampler from Section 4.3.2 is run 25 times with the same seed, but different starting values. This is done to find out whether, depending on the starting value, different sparse structures are found, as described in Section 4.3.4.

Table 4.14 shows relative frequencies for different numbers of factors identified by the two-step approach. For $\alpha = 0.1$, four factors are found in 75% of the cases, and five factors are found in 25% of the cases. For $\alpha = 0.05$, the relative frequency of five-factor outcomes is reduced to 5%, and for $\alpha = 0.01$, four factors are found in all cases. The sparse sampler chooses a model with five factors in 80% of the cases, and prefers four factors in 20% of the cases. The lower part of the table shows that the two-step approach does not identify any irrelevant variables, independent of the chosen α. The same holds for the sparse sampler. Table 4.15 shows the relative frequencies for the total number of zero elements in Λ identified by the sparse approach with $\alpha = 0.1$, $\alpha = 0.05$ and $\alpha = 0.01$, and by the sparse sampler from Section 4.3.2. The number of nonzero loadings postulated by Holzinger and Swineford (1939) is 48, which is larger than the results obtained from the sparse sampler obtained with $\alpha = 0.05$ and $\alpha = 0.01$. The sparse sampler finds a structure with 51 nonzero elements in Λ in 44% of the cases, but also finds many other degrees of sparsity. Altogether, this indicates that multimodality might be an issue here: 51 nonzero elements is the most frequently found degree of sparsity, but not the most parsimonious structure. The variation in the number of nonzero elements generally seems to be smaller in the two-step approach, where the "lock-in" effect of the sparse sampler is absent. Regarding the sparse sampler, however, convergence properties, prior sensitivity and the effect of choosing the median as the estimator should be checked to find explanations for the different behavior.

Figure 4.13 shows the association probabilities for each element of Λ obtained from the two-step approach for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Since the association probability can only be zero or one, the plots show the average over the 20 runs of the sampler. The black frames

[20] Assuming that the nonlinear optimization step always succeeds to find a unique globally most parsimonious structure in Λ, differences in the structure could then be attributed to Monte Carlo variation. If there does not exist a globally unique most parsimonious structure, repeated estimations increase the chance of finding all structures with the same degree of sparsity. Eventually, if the nonlinear optimization step does not always find a most parsimonious structure, it may instead find local modes in some of the 20 instances.

indicate where the bi-factor structure by Holzinger and Swineford (1939) assumes the nonzero elements of Λ. Ideally, if the results were identical in repeated estimations, the association probabilities would all be zero or one. Those $\lambda_{i,k}$ that were found to be zero in all 20 runs of the sampler are indicated as grey patches. Conversely, the $\lambda_{i,k}$ that were found to be nonzero in all 20 runs are indicated as red patches. The lighter red to white patches have association probabilities which are positive, but less than one. For $\alpha = 0.01$, 37 out of the 51 different loadings that are included in the model in at least one estimate are always included. For $\alpha = 0.05$, this holds only for 25 out of 67 loadings, and for $\alpha = 0.1$, it holds only for 34 out of 73 loadings included in at least one estimate. The latter result does not come unexpected if it is taken into account that the number of factors varies over the 20 results obtained in the latter two cases. For $\alpha = 0.01$ and $\alpha = 0.05$, the fifth group factor, the "mathematical deduction" factor, which Holzinger and Swineford (1939) drop from the analysis, is not found in any of the 20 runs of the sampler, whereas it is found at least once for $\alpha = 0.1$. The first group factor, the "spatial" factor is not found in any of the 20 runs for $\alpha = 0.01$.

Similarly, Figure 4.14 shows the association probabilities determined from the sparse sampler. They can take any value between zero and one in each iteration of the sampler, so the posterior estimate for the association probability likewise takes a value between zero and one for each $\lambda_{i,k}$. The plot shows the 16%, the 50% and the 84% quantiles over the estimates from the 25 runs of the sampler. The quantiles are chosen to provide an impression of the distribution of association probabilities. Ideally, the estimated association probabilities would be the same for all 25 runs. Conversely, a wide range of estimates for the association probabilities indicates that the sparse sampler converges to different sparse patterns, depending on the starting value, as shown in Section 4.3.4.

Figure 4.15 shows the average over the loadings estimates from the two-step approach for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. The grey patches, in line with the grey patches in Figure 4.13, indicate values that are always zero. White patches indicate that in almost all out of the 20 runs, the corresponding loadings have been set to zero. For comparison, Figure 4.16 shows the average loadings found from the PC factor analysis and from the WOP approach, where for both, the results have been rotated to minimize the distance to the postulated bi-factor structure, and the average loadings found from the sparse sampler. The loadings estimates that are part of the bi-factor structure generally look similar to the loadings estimates from the two-step approach. The estimates from the sparse sampler show that in 25 runs, the fifth group factor is found in none of the 25 runs of the sampler. Both the sparse sampler and the two-step approach for different choices of α yields a negative loading of the variable "Flags" on the second group factor, the "verbal" factor, whereas all the other loadings have the expected positive sign. The variable "Add" is found not to load on the general factor by the two-step approach for $\alpha = 0.01$ and by the sparse sampler, which additionally finds the variable "Object-Number" not to load on the general factor. The variable "Arithmetic Problems" is found to load on multiple additional factors occasionally both in the sparse and the two-step approach. Similarly, in both approaches, the variables "Problem Reasoning" and "Code" are found to load on the second group factor, the "verbal" factor, and the variables

"Object-Number", "Number-Figure", "Numerical Puzzles" and "Arithmetic Problems" are found to load on the third group factor, the "speed" factor. Figure 4.17 shows the scatter plots for the sum of squared loadings for each pair of factors, except for those pairs including the fifth group factor, where all loadings are zero. The blue circles denote the results from the sparse sampler, and the red 'x' marks denote the results from the two-step approach with $\alpha = 0.01$. For both samplers, clearly not all results are identical, and, in particular for the general factor, the loadings appear to be larger in magnitude in the estimates from the two-step approach.

Overall, the estimated factor loadings from the two-step approach are similar to each other for all three different choices of α, although increasing the level of α yields additional nonzero loadings. On a side note, the numerical standard deviations calculated for the $\lambda_{i,k}$ found to be nonzero in all 20 runs of the sampler are very small, i.e. a change in the other elements of Λ does not appear to affect them much. The estimates from the sparse approach resemble those from the two-step approach, although they are generally slightly smaller in magnitude on average. For the first group factor, the "spatial" factor, they are larger in magnitude on average, which is due to the fact that the corresponding loadings are set to zero less often than in the two-step approach. Both the results from the sparse sampler and the two-step approach find a sparse structure which is similar to the postulated bi-factor structure, with the exception of the fifth group factor, which is in line with the findings of Holzinger and Swineford (1939). The set of variables for which additional nonzero loadings on some of the factors are found are overall similar for both approaches.

4.7 Conclusion

In this chapter, I propose a two-step approach to find a parsimonious loadings structure for static factor models. The approach is based on the weighted orthogonal Procrustes (WOP) procedure, whose outcome is used to construct multivariate HPD intervals, or HPD ellipsoids (HPDEs) for the factor loadings Λ, along the lines of an approach suggested by Hanson and McMillan (2012) and uses the properties of the posterior densities obtained from the WOP approach to find an orthogonal transformation of these HPDEs to have them include the zero for as many elements of the loadings matrix as possible. The orthogonal transformation is found by means of a nonlinear optimization using a Simplex-type algorithm along the lines of Nelder and Mead (1965). The approach can likewise be used to set entire rows from the loadings matrix to zero, thus identifying variables irrelevant for the analysis or, in line with the identification requirements stated by Anderson and Rubin (1956), to reduce the number of nonzero elements per column to less than three, thus reducing the number of factors.

In a simulation study, I generate 50 data sets following a specification used in Lopes and West (2004) and Frühwirth-Schnatter and Lopes (2012) to evaluate how well the two-step approach recovers a given sparse structure in a loadings matrix. Choosing different values for α, which governs the widths of the constructed HPDEs, I find that the structure can be recovered

about as well as with the sampling approach by Carvalho (2006) in the version of Kaufmann and Schumacher (2012). The root-mean squared errors, however, are smaller in the two-step approach. In a second simulation study, I add three irrelevant variables, consisting only of noise, which are also successfully identified by the approach in 50 generated data sets. In a third simulation study, I estimate the factor model with the number of factors specified too large. The two-step approach succeeds in eliminating the additional factor in all cases. The findings from the first experiment when comparing the two-step approach to the sparse sampler are confirmed in the second and third experiment.

Eventually, I analyze a data set by Holzinger and Swineford (1939) initially used to establish the bi-factor model, a structure in which each variable loads on the same general factor and an additional group-specific factor. Without any prior knowledge about the number of factors in the model, the relevance of the variables in the data set and the loadings structure, I use the two-step approach to answer these three questions. I find that, depending on the width of the HPDEs, the approach determines the number of factors to be either four or five. The bi-factor structure is overall confirmed by the two-step approach, the deviations that are found are generally comprehensible. Running the sparse sampler for comparison yields results similar to those from the two-step approach.

Tables

factor founder	1	2	3	4	5
minimum	-4717.59	-4731.08	-4402.57	-3875.11	-3923.95
angle	0.000	1.553	1.065	0.663	0.768
factor founder	6	7	8	9	10
minimum	-4563.52	-4839.93	-4506.80	-3845.41	-3695.49
angle	1.257	0.890	0.593	1.065	0.192

Table 4.1: Minimum of the log likelihood under different founders for the first factor and location of the minimum in the first quadrant.

permutation	1	2	3	4	5
maximum	-6079.61	-6094.72	-6456.30	-6117.11	-6464.34
minimum	-6276.87	-6553.96	-6514.96	-6275.78	-6485.49
angle max	1.327	1.431	0.052 1.047	1.292	0.140
angle min	1.396	1.030	0.314 0.768	1.414	1.518
permutation	6	7	8	9	10
maximum	-6108.92	-6253.18	-6319.95	-6228.67	-6258.78
minimum	-6393.20	-6514.06	-6534.55	-6506.43	-6523.95
angle max	1.152	0.977	0.035 0.611	1.012	1.030
angle min	0.925	0.873	1.065 1.100	0.890	0.873

Table 4.2: Minimum and maximum of the log likelihood under different row permutations and location of the minima and maxima in the first quadrant.

Nonzero elements	90% HPDEs	95% HPDEs	99% HPDEs	sparse sampler
9	0.96	0.98	1.00	0.98
10	0.02	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00
12	0.02	0.02	0.00	0.02

Table 4.3: Experiment 1: Number of nonzero elements identified by the two-step procedure.

Notes: The table reports the relative frequency of each outcome for three different values of α in creating the HPDEs in the first three columns, and for the sparse sampler described in Section 4.3.2 in the last column.

$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.1$		
0.9882 (0.0072)	0 (0)	0 (0)	0.9875 (0.0057)	0 (0)	0 (0)	0.9876 (0.0115)	0 (0)	0 (0)
0 (0)	0.9100 (0.0262)	0 (0)	0 (0)	0.9069 (0.0238)	−0.0053 (0.0371)	0 (0)	0.9087 (0.0259)	0 (0)
0 (0)	0 (0)	0.8308 (0.0441)	0 (0)	0 (0)	0.8228 (0.0569)	0.0015 (0.0108)	0 (0)	0.8341 (0.0496)
0.9882 (0.0070)	0 (0)	0 (0)	0.9874 (0.0060)	0 (0)	0 (0)	0.9878 (0.0108)	0 (0)	−0.0067 (0.0469)
0.9885 (0.0071)	0 (0)	0 (0)	0.9877 (0.0058)	0 (0)	0 (0)	0.9876 (0.0110)	0 (0)	−0.0068 (0.0476)
0 (0)	0.9118 (0.0199)	0 (0)	0 (0)	0.9100 (0.0218)	−0.0086 (0.0604)	0 (0)	0.9076 (0.0202)	0 (0)
0 (0)	0.9085 (0.0241)	0 (0)	0 (0)	0.9090 (0.0277)	−0.0069 (0.0484)	0 (0)	0.9099 (0.0182)	0 (0)
0 (0)	0 (0)	0.8193 (0.0396)	0 (0)	0 (0)	0.8181 (0.0428)	0 (0)	0 (0)	0.8357 (0.0368)
0 (0)	0 (0)	0.8255 (0.0398)	0 (0)	0 (0)	0.8293 (0.0490)	0 (0)	0 (0)	0.8271 (0.0477)

Table 4.4: Experiment 1: Mean estimates for the sparse structure found by the two-step approach with $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Standard deviations over the 50 estimates in parentheses.

PC and Varimax			WOP			sparse sampler		
0.9903 (0.0029)	−0.0042 (0.0548)	0.0137 (0.0537)	1.0304 (0.0061)	0.0016 (0.0174)	0.0016 (0.0151)	0.9146 (0.0077)	0.0047 (0.0332)	0 (0)
0.0028 (0.0545)	0.9389 (0.0135)	−0.0103 (0.0659)	−0.0041 (0.0959)	0.9380 (0.0271)	−0.0124 (0.0771)	0 (0)	0.8474 (0.0201)	0 (0)
0.0043 (0.0699)	−0.0124 (0.0766)	0.8869 (0.0212)	0.0180 (0.1073)	−0.0122 (0.0963)	0.8520 (0.0477)	0 (0)	0 (0)	0.7782 (0.0433)
0.9906 (0.0023)	−0.0078 (0.0518)	0.0142 (0.0517)	1.0305 (0.0058)	−0.0022 (0.0170)	0.0020 (0.0169)	0.9148 (0.0068)	0.0041 (0.0288)	0 (0)
0.9908 (0.0022)	−0.0029 (0.0514)	0.0109 (0.0520)	1.0307 (0.0057)	0.0026 (0.0176)	−0.0012 (0.0165)	0.9150 (0.0075)	0.0046 (0.0325)	0 (0)
−0.0115 (0.0626)	0.9388 (0.0105)	−0.0042 (0.0627)	−0.0170 (0.1051)	0.9385 (0.0222)	−0.0069 (0.0636)	0 (0)	0.8396 (0.0197)	0 (0)
−0.0059 (0.0616)	0.9374 (0.0129)	−0.0086 (0.0611)	−0.0115 (0.1030)	0.9352 (0.0265)	−0.0099 (0.0627)	0 (0)	0.8411 (0.0219)	0 (0)
0.0078 (0.0635)	−0.0028 (0.0754)	0.8836 (0.0202)	0.0203 (0.1031)	−0.0040 (0.0890)	0.8431 (0.0421)	0 (0)	0 (0)	0.7719 (0.0443)
0.0229 (0.0656)	−0.0080 (0.0798)	0.8853 (0.0198)	0.0336 (0.1053)	−0.0081 (0.0954)	0.8473 (0.0426)	0 (0)	0 (0)	0.7712 (0.0382)

Table 4.5: Experiment 1: Mean estimates for the sparse structure found by PC factor analysis with Varimax rotation, by the WOP approach and by the sparse sampler described in Section 4.3.2. Standard deviations over the 50 estimates in parentheses.

Rows with nonzero elements	90% HPDEs	95% HPDEs	99% HPDEs	sparse sampler
9	0.98	1.00	1.00	0.92
10	0.02	0.00	0.00	0.08
11	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00
Nonzero elements	**90% HPDEs**	**95% HPDEs**	**99% HPDEs**	**sparse sampler**
9	0.86	0.96	1.00	0.92
10	0.08	0.02	0.00	0.08
11	0.02	0.02	0.00	0.00
12	0.04	0.00	0.00	0.00

Table 4.6: Experiment 2: Number of nonzero rows and nonzero elements identified by the two-step procedure for the data set with three additional rows of zeros.

Notes: The table reports the relative frequency of each outcome for three different values of α in creating the HPDEs in the first three columns, and for the sparse sampler described in Section 4.3.2 in the last column.

$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.1$		
0.9876 (0.0060)	0 (0)	0 (0)	0.9882 (0.0071)	0 (0)	0 (0)	0.9870 (0.0084)	−0.0048 (0.0301)	0.0044 (0.0307)
0 (0)	0.9146 (0.0228)	0 (0)	0 (0)	0.9056 (0.0254)	0 (0)	−0.0023 (0.0163)	0.9085 (0.0231)	0 (0)
0 (0)	0 (0)	0.8275 (0.0554)	0.0025 (0.0177)	0 (0)	0.8294 (0.0534)	0 (0)	0 (0)	0.8310 (0.0384)
0.9879 (0.0063)	0 (0)	0 (0)	0.9881 (0.0074)	0 (0)	0 (0)	0.9870 (0.0083)	−0.0041 (0.0288)	0.0043 (0.0298)
0.9880 (0.0062)	0 (0)	0 (0)	0.9883 (0.0072)	0 (0)	0 (0)	0.9871 (0.0082)	−0.0039 (0.0275)	0.0042 (0.0294)
0 (0)	0.9146 (0.0180)	0 (0)	0 (0)	0.9022 (0.0278)	0 (0)	0 (0)	0.9107 (0.0207)	−0.0024 (0.0171)
0 (0)	0.9118 (0.0206)	0 (0)	0 (0)	0.9155 (0.0220)	0 (0)	0 (0)	0.9069 (0.0237)	0.0025 (0.0174)
0 (0)	0 (0)	0.8256 (0.0469)	0 (0)	0 (0)	0.8226 (0.0519)	0 (0)	0 (0)	0.8436 (0.0348)
0 (0)	0 (0)	0.8159 (0.0484)	−0.0010 (0.0379)	0 (0)	0.8252 (0.0509)	0 (0)	−0.0043 (0.0301)	0.8354 (0.0418)
0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.0067 (0.0466)
0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Table 4.7: Experiment 2: Mean estimates for the sparse structure found by the two-step approach with $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Standard deviations over the 50 estimates in parentheses.

PC and Varimax			WOP			sparse sampler		
0.9868 (0.0036)	−0.0098 (0.0462)	0.0016 (0.0503)	1.0468 (0.0081)	0.0004 (0.0181)	−0.0027 (0.0161)	0.9135 (0.0038)	0 (0)	0 (0)
0.0029 (0.0664)	0.9354 (0.0151)	−0.0023 (0.0684)	−0.0103 (0.0860)	0.9586 (0.0236)	−0.0074 (0.0604)	0 (0)	0.8458 (0.0157)	0 (0)
−0.0063 (0.0840)	0.0131 (0.0699)	0.8783 (0.0277)	−0.0030 (0.0994)	0.0174 (0.0844)	0.8650 (0.0582)	0 (0)	0 (0)	0.7786 (0.0372)
0.9869 (0.0040)	−0.0106 (0.0453)	0.0064 (0.0518)	1.0471 (0.0085)	−0.0004 (0.0154)	0.0029 (0.0122)	0.9132 (0.0036)	0 (0)	0 (0)
0.9871 (0.0035)	−0.0067 (0.0484)	0.0011 (0.0528)	1.0472 (0.0082)	0.0038 (0.0187)	−0.0025 (0.0122)	0.9131 (0.0038)	0 (0)	0 (0)
0.0100 (0.0638)	0.9355 (0.0109)	0.0052 (0.0605)	−0.0025 (0.0867)	0.9593 (0.0199)	0.0008 (0.0530)	0 (0)	0.8440 (0.0217)	0 (0)
0.0030 (0.0613)	0.9353 (0.0142)	0.0004 (0.0709)	−0.0105 (0.0847)	0.9563 (0.0222)	−0.0041 (0.0676)	0 (0)	0.8421 (0.0207)	0 (0)
−0.0088 (0.0860)	−0.0014 (0.0753)	0.8743 (0.0292)	−0.0049 (0.1053)	0.0018 (0.0882)	0.8594 (0.0512)	0 (0)	0 (0)	0.7731 (0.0393)
0.0059 (0.0907)	0.0014 (0.0788)	0.8705 (0.0278)	0.0074 (0.1077)	0.0065 (0.0921)	0.8499 (0.0506)	0 (0)	0 (0)	0.7808 (0.0465)
−0.0433 (0.1699)	−0.0123 (0.1438)	0.0083 (0.1876)	−0.0313 (0.1241)	−0.0069 (0.1017)	0.0064 (0.1266)	−0.0051 (0.0357)	0 (0)	0 (0)
−0.0190 (0.1297)	0.0113 (0.1480)	0.0194 (0.1645)	−0.0121 (0.0908)	0.0037 (0.1017)	0.0144 (0.1118)	0 (0)	0.0043 (0.0300)	0 (0)
0.0288 (0.1470)	0.0331 (0.1714)	0.0263 (0.1890)	0.0194 (0.1096)	0.0238 (0.1182)	0.0164 (0.1284)	−0.0045 (0.0316)	0.0052 (0.0362)	0 (0)

Table 4.8: Experiment 2: Mean estimates for the sparse structure found by PC factor analysis with Varimax rotation, by the WOP approach and by the sparse sampler described in Section 4.3.2. Standard deviations over the 50 estimates in parentheses.

Number of factors	90% HPDEs	95% HPDEs	99% HPDEs	sparse sampler
3	1.00	1.00	1.00	1.00
4	0.00	0.00	0.00	0.00
Nonzero elements	**90% HPDEs**	**95% HPDEs**	**99% HPDEs**	**sparse sampler**
9	0.94	0.96	1.00	0.98
10	0.02	0.04	0.00	0.00
11	0.04	0.00	0.00	0.00
12	0.00	0.00	0.00	0.02

Table 4.9: Experiment 3: Number of factors and nonzero elements identified by the two-step procedure for the data set, starting with $K_{max} = 4$.

Notes: The table reports the relative frequency of each outcome for three different values for α in creating the HPDEs in the first three columns, and for the sparse sampler described in Section 4.3.2 in the last column.

$\alpha = 0.01$				$\alpha = 0.05$				$\alpha = 0.1$			
0.9889 (0.0076)	0 (0)	0 (0)	0 (0)	0.9899 (0.0066)	0 (0)	0 (0)	0 (0)	0.9878 (0.0066)	0 (0)	0 (0)	0 (0)
0 (0)	0.9057 (0.0236)	0 (0)	0 (0)	0 (0)	0.9126 (0.0177)	0 (0)	0 (0)	0 (0)	0.9113 (0.0184)	0 (0)	0 (0)
0 (0)	0 (0)	0.8320 (0.0374)	0 (0)	0 (0)	0.0010 (0.0424)	0.8264 (0.0457)	0 (0)	0 (0)	0 (0)	0.8328 (0.0432)	0 (0)
0.9894 (0.0081)	0 (0)	0 (0)	0 (0)	0.9897 (0.0067)	0 (0)	0 (0)	0 (0)	0.9877 (0.0061)	0 (0)	0 (0)	0 (0)
0.9892 (0.0078)	0 (0)	0 (0)	0 (0)	0.9900 (0.0071)	0 (0)	0 (0)	0 (0)	0.9881 (0.0066)	0 (0)	0 (0)	0 (0)
0 (0)	0.9069 (0.0209)	0 (0)	0 (0)	0 (0)	0.9059 (0.0173)	0 (0)	0 (0)	0 (0)	0.9085 (0.0209)	0 (0)	0 (0)
0.0019 (0.0135)	0.9085 (0.0174)	0 (0)	0 (0)	0 (0)	0.9078 (0.0236)	0 (0)	0 (0)	0 (0)	0.9095 (0.0224)	0 (0)	0 (0)
0.0026 (0.0180)	0 (0)	0.8210 (0.0476)	0 (0)	−0.0035 (0.0246)	0.0059 (0.0414)	0.8220 (0.0401)	0 (0)	0 (0)	0 (0)	0.8391 (0.0371)	0 (0)
0 (0)	0 (0)	0.8359 (0.0410)	0 (0)	−0.0028 (0.0198)	0 (0)	0.8303 (0.0416)	0 (0)	0 (0)	0 (0)	0.8283 (0.0394)	0 (0)

Table 4.10: Experiment 3: Mean estimates for the sparse structure found by the two-step approach with $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Standard deviations over the 50 estimates in parentheses.

PC and Varimax				WOP				sparse sampler			
0.9910 (0.0018)	0.0021 (0.0436)	−0.0044 (0.0431)	0.0018 (0.0253)	1.0325 (0.0161)	0.0020 (0.0161)	0.0011 (0.0382)	0.0098 (0.0644)	0.9133 (0.0074)	0 (0)	0.0050 (0.0349)	0 (0)
−0.0039 (0.0508)	0.9403 (0.0104)	−0.0004 (0.0466)	−0.0058 (0.0763)	−0.0045 (0.0882)	0.9444 (0.0234)	0.0061 (0.0760)	−0.0013 (0.1011)	0 (0)	0.8415 (0.0208)	0 (0)	0 (0)
−0.0084 (0.0713)	0.0215 (0.0683)	0.8471 (0.1278)	−0.0329 (0.4037)	−0.0171 (0.1103)	0.0257 (0.1049)	0.8423 (0.1013)	0.0501 (0.1903)	0 (0)	0 (0)	0.7763 (0.0420)	0 (0)
0.9909 (0.0021)	−0.0004 (0.0460)	−0.0069 (0.0437)	0.0007 (0.0221)	1.0322 (0.0171)	−0.0010 (0.0189)	−0.0011 (0.0372)	0.0106 (0.0675)	0.9131 (0.0083)	0 (0)	0.0052 (0.0366)	0 (0)
0.9910 (0.0020)	0.0006 (0.0477)	−0.0050 (0.0449)	−0.0043 (0.0211)	1.0328 (0.0162)	0.0001 (0.0197)	0 (0.0318)	0.0075 (0.0655)	0.9131 (0.0072)	0 (0)	0.0049 (0.0342)	0 (0)
−0.0061 (0.0595)	0.9378 (0.0115)	0.0252 (0.0556)	−0.0042 (0.0810)	−0.0071 (0.0950)	0.9399 (0.0263)	0.0226 (0.0697)	0.0080 (0.1138)	0 (0)	0.8415 (0.0228)	0 (0)	0 (0)
0.0118 (0.0571)	0.9380 (0.0113)	0.0082 (0.0647)	0.0028 (0.0757)	0.0116 (0.0934)	0.9434 (0.0216)	0.0099 (0.0710)	0.0145 (0.0848)	0 (0)	0.8440 (0.0193)	0 (0)	0 (0)
0.0010 (0.0526)	0.0062 (0.0649)	0.8474 (0.1348)	−0.0866 (0.3908)	−0.0049 (0.0880)	0.0093 (0.0963)	0.8498 (0.1294)	0.0044 (0.1835)	0 (0)	0 (0)	0.7623 (0.0554)	0 (0)
−0.0009 (0.0674)	0.0119 (0.0585)	0.8155 (0.1355)	−0.0915 (0.4633)	−0.0064 (0.1055)	0.0141 (0.1056)	0.8412 (0.0930)	0.0075 (0.1848)	0 (0)	0 (0)	0.7745 (0.0366)	0 (0)

Table 4.11: Experiment 3 : Mean estimates for the sparse structure found by PC factor analysis with Varimax rotation, by the WOP approach and by the sparse sampler described in Section 4.3.2. Standard deviations over the 50 estimates in parentheses.

	Λ					Σ				
	$\xi_{(1)}(\lambda)$	$\xi_{(3)}(\lambda)$	$\xi_{(5)}(\lambda)$	$\xi_{(7)}(\lambda)$	$\xi_{(9)}(\lambda)$	$\xi_{(1)}(\sigma^2)$	$\xi_{(3)}(\sigma^2)$	$\xi_{(5)}(\sigma^2)$	$\xi_{(7)}(\sigma^2)$	$\xi_{(9)}(\sigma^2)$
Experiment 1 : Finding a parsimonious structure										
PCA & Varimax	0.0034	0.0036	0.0214	0.0246	0.0287	0.0067	0.0067	0.0746	0.1564	0.1615
WOP	0.0034	0.0036	0.0214	0.0246	0.0287	0.0067	0.0067	0.0746	0.1564	0.1615
Sparse sampler	0.0754	0.0758	0.1111	0.1293	0.1356	0.0153	0.0162	0.0369	0.0699	0.0832
Two-step approach	0.0112	0.0126	0.0506	0.0838	0.0911	0.0159	0.0168	0.0374	0.0693	0.0825
Experiment 2 : Identifying irrelevant variables										
PCA & Varimax	0.0052	0.0057	0.0214	0.0313	0.0362	0.0069	0.0070	0.0720	0.1489	0.1561
WOP	0.0052	0.0057	0.0214	0.0313	0.0362	0.0069	0.0070	0.0720	0.1489	0.1561
Sparse sampler	0.0766	0.0770	0.1081	0.1270	0.1328	0.0150	0.0157	0.0386	0.0738	0.0904
Two-step approach	0.0087	0.0092	0.0479	0.0800	0.0870	0.0156	0.0163	0.0392	0.0736	0.0896
Experiment 3 : Determining the number of factors										
PCA & Varimax	0.0030	0.0032	0.0179	0.0242	0.0339	0.0068	0.0072	0.0752	0.1520	0.1584
WOP	0.0029	0.0031	0.0179	0.1387	0.1992	0.0071	0.0075	0.0799	0.2651	0.2869
Sparse sampler	0.0771	0.0774	0.1104	0.1307	0.1484	0.0153	0.0155	0.0384	0.0686	0.0864
Two-step approach	0.0068	0.0069	0.0505	0.0836	0.1045	0.0158	0.0161	0.0392	0.0682	0.0876

Table 4.12: Root mean-squared errors for the nonzero elements of Λ and the diagonal elements of Σ for the three estimation procedures for all three experiments.

Notes: For the two-step approach, $\alpha = 0.05$ is chosen. Instead of reporting the RMSE for all nine parameters each, the RMSEs are ordered, and only the first, third, fifth, seventh, and ninth are reported.

1	Visual Perception	13	Straight and Curved Capitals
2	Cubes	14	Word Recognition
3	Paper Form Board	15	Number Recognition
4	Flags	16	Figure Recognition
5	General Information	17	Object-Number
6	Paragraph Comprehension	18	Number-Figure
7	Sentence Completion	19	Figure-Word
8	Word Classification	20	Deduction
9	Word Meaning	21	Numerical Puzzles
10	Add	22	Problem Reasoning
11	Code	23	Series Completion
12	Counting Groups of Dots	24	Arithmetic Problems

Table 4.13: List of the 24 tasks from the original study by Holzinger and Swineford (1939).

Number of factors	90% HPDEs	95% HPDEs	99% HPDEs	sparse sampler
4	0.75	0.95	1.00	0.20
5	0.25	0.05	0.00	0.80
Rows with nonzero elements	**90% HPDEs**	**95% HPDEs**	**99% HPDEs**	
24	1.00	1.00	1.00	1.00

Table 4.14: Number of factors, nonzero rows and nonzero elements identified by the two-step procedure for the data set from Holzinger and Swineford (1939).

Notes: The table reports the relative frequency of each outcome for three different values for α in the HPDEs in the first three columns and for the sparse sampler from Section 4.3.2 in the last column.

Nonzero elements	90% HPDEs	95% HPDEs	99% HPDEs	sparse sampler
40	0.00	0.00	0.35	0.00
41	0.00	0.00	0.25	0.00
42	0.00	0.00	0.35	0.00
43	0.00	0.30	0.00	0.00
44	0.00	0.50	0.05	0.00
45	0.05	0.15	0.00	0.12
46	0.35	0.05	0.00	0.12
47	0.25	0.00	0.00	0.08
48	0.25	0.00	0.00	0.04
49	0.00	0.00	0.00	0.12
50	0.05	0.00	0.00	0.04
51	0.05	0.00	0.00	0.44
52	0.00	0.00	0.00	0.04

Table 4.15: Number of nonzero elements identified by the two-step procedure for the data set from Holzinger and Swineford (1939).

Notes: The table reports the relative frequency of each outcome for three different values for α in the HPDEs in the first three columns and for the sparse sampler from Section 4.3.2 in the last column.

Figures

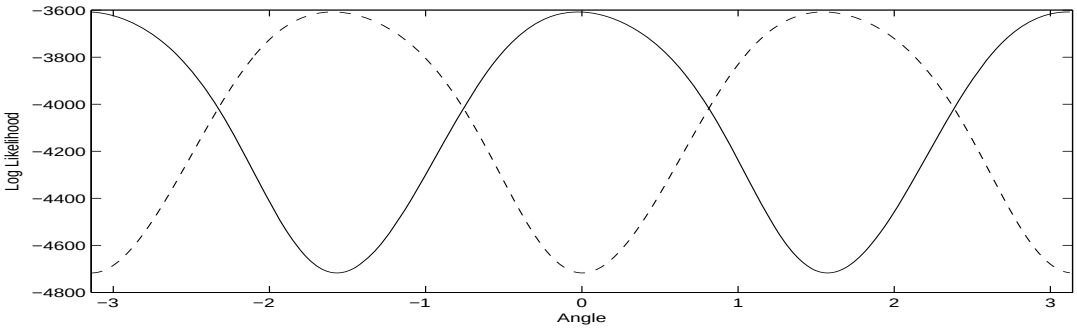

Figure 4.1: Modes of the log likelihood.

Notes: Solid line shows the rotated solution, dashed line shows the solutions with columns exchanged and rotated.

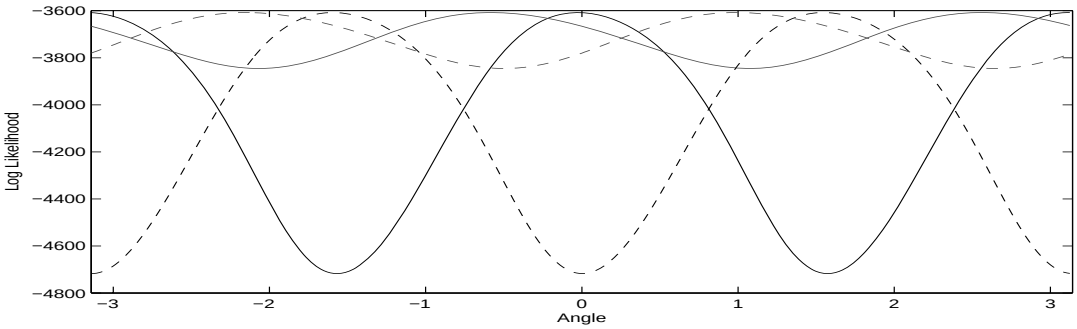

Figure 4.2: Modes of the log likelihood with alternative model identification.

Notes: Solid line shows the rotated solution, dashed line shows the solutions with columns exchanged and rotated. Second line shows an alternative model identification.

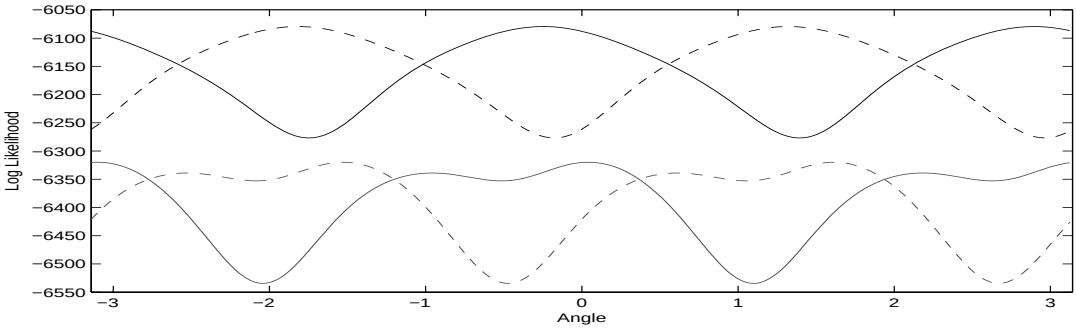

Figure 4.3: Modes of the log likelihood with a sparse loadings matrix.

Notes: Solid line shows the rotated solution, dashed line shows the solutions with columns exchanged and rotated. Second line shows a sparsity pattern with permuted rows of Δ.

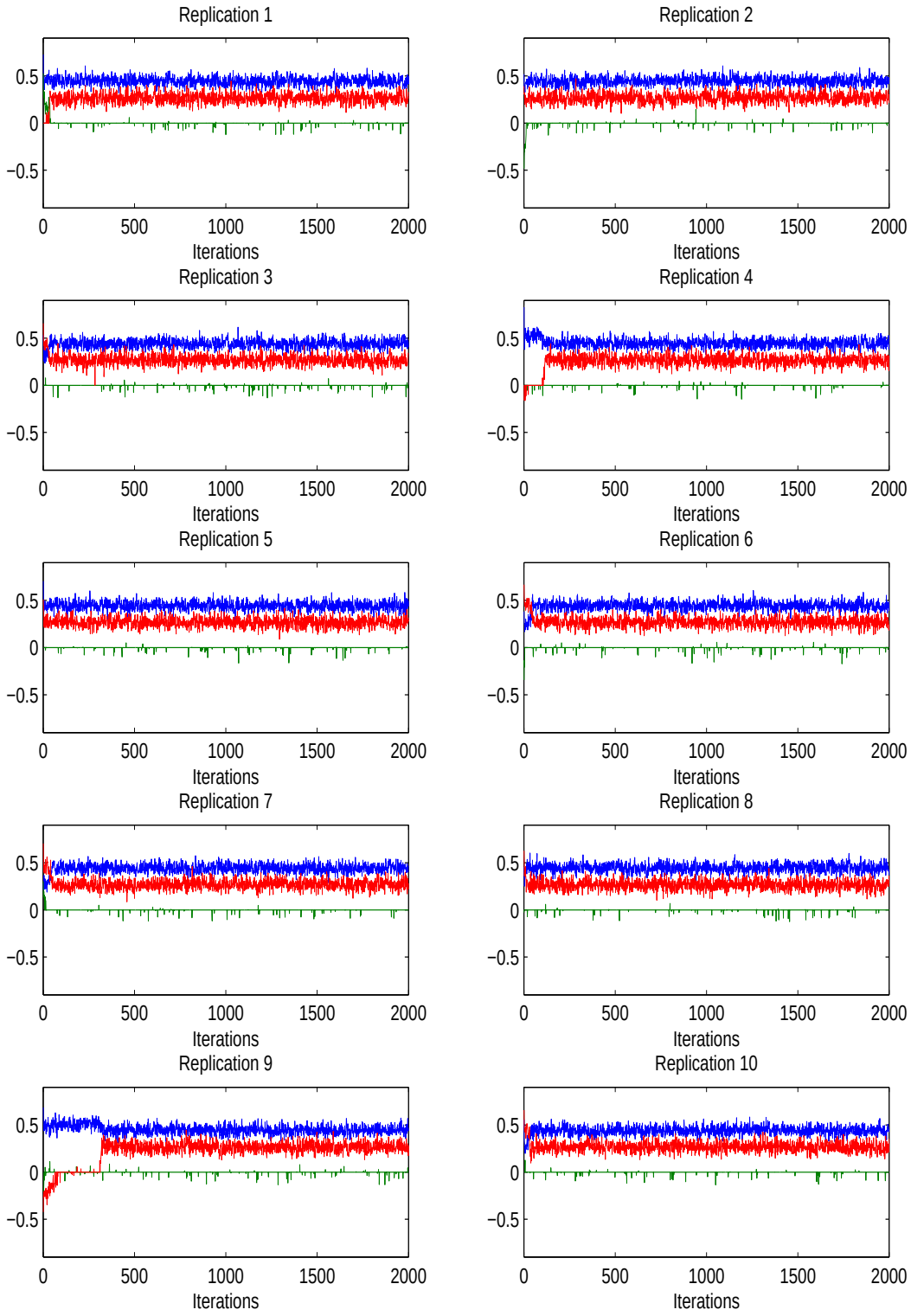

Figure 4.4: First 2,000 iterations from 10 sequences of the factor loadings of one randomly selected variable from the sampler of Kaufmann and Schumacher (2013) for simulated data with 73% zero elements in Λ, using different starting points.

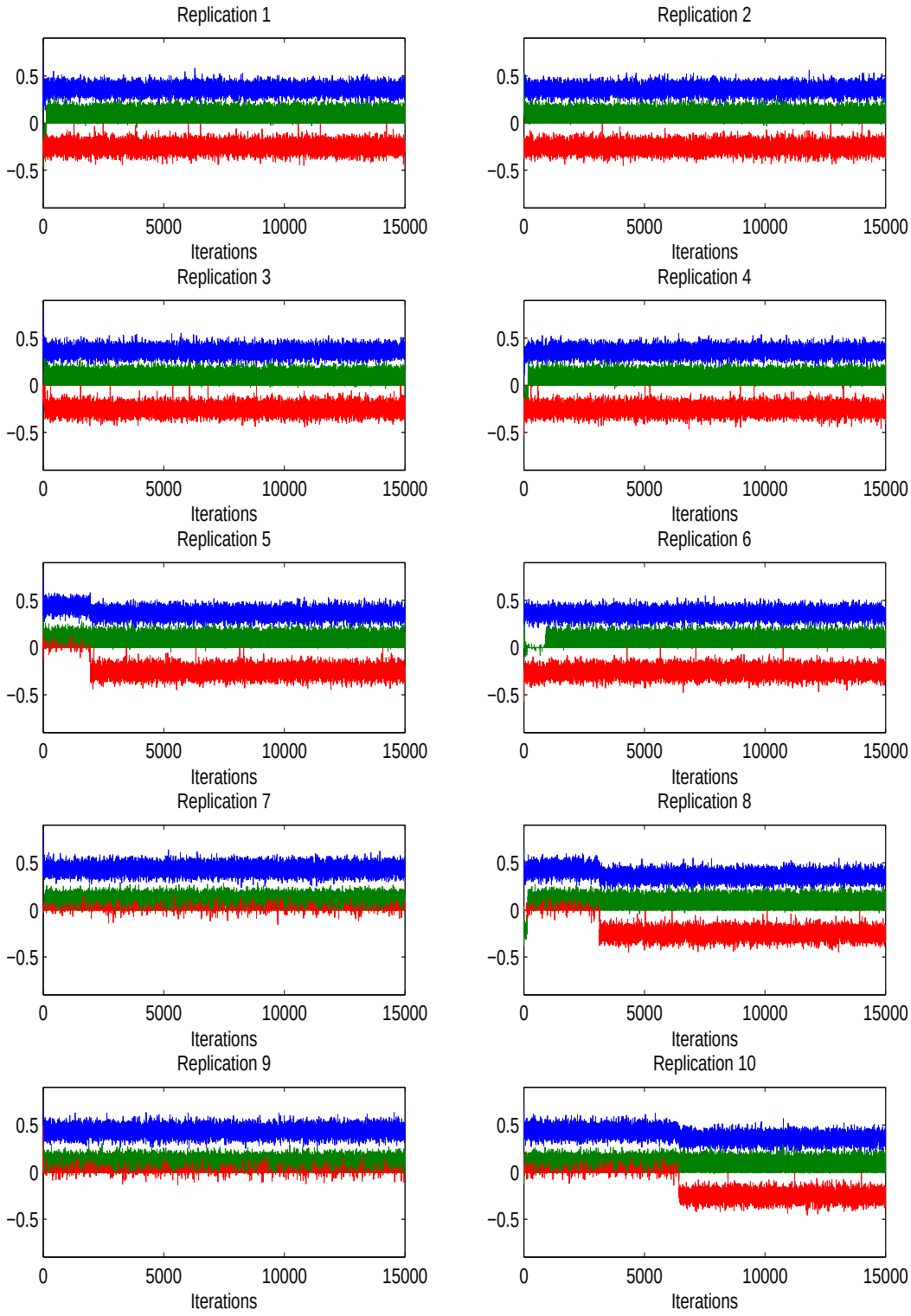

Figure 4.5: First 15,000 iterations from 10 sequences of the factor loadings of one randomly selected variable from the sampler of Kaufmann and Schumacher (2013) for simulated data with 55% zero elements in Λ, using different starting points.

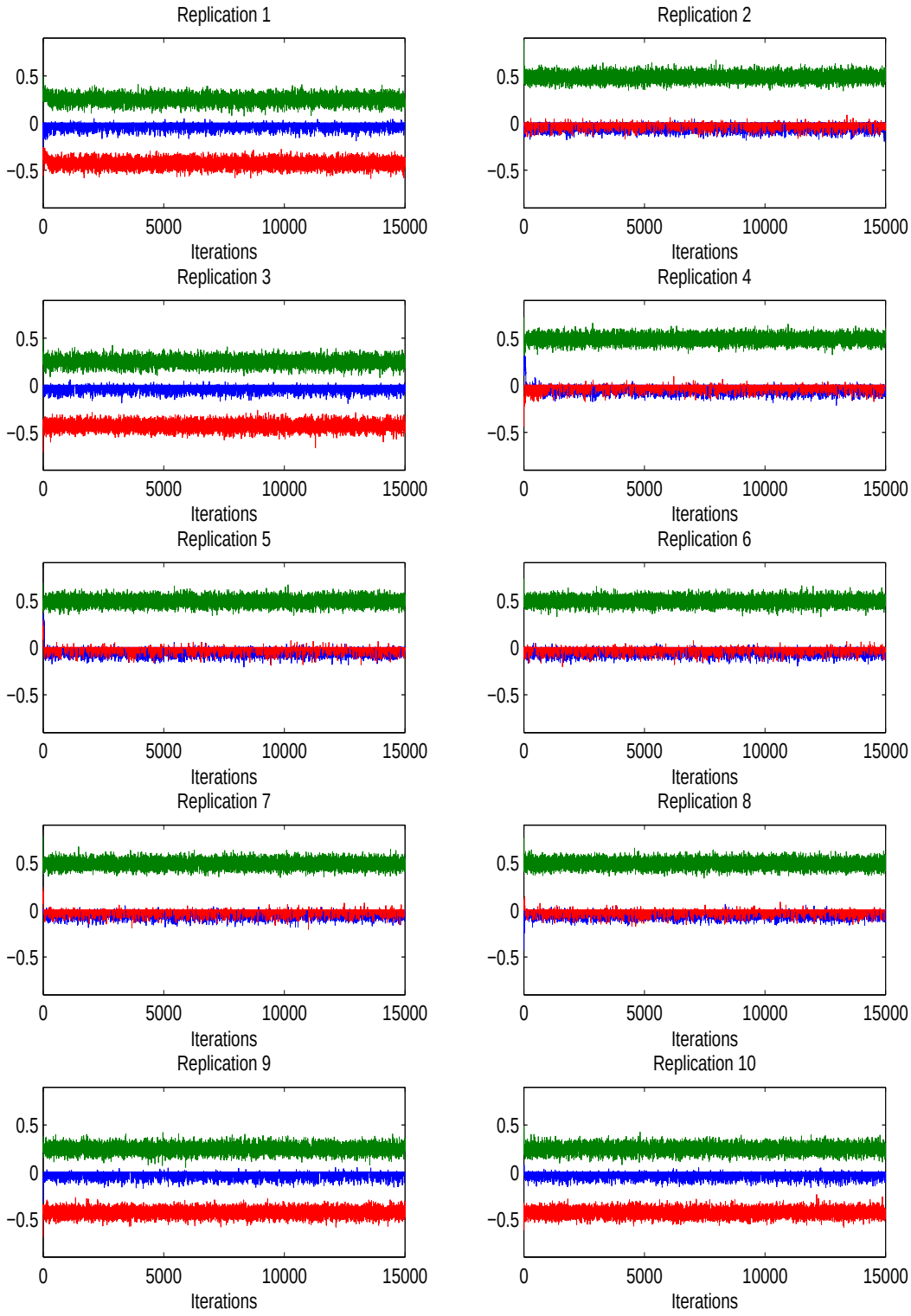

Figure 4.6: First 15,000 iterations from 10 sequences of the factor loadings of one randomly selected variable from the sampler of Kaufmann and Schumacher (2013) for simulated data with 34% zero elements in Λ, using different starting points.

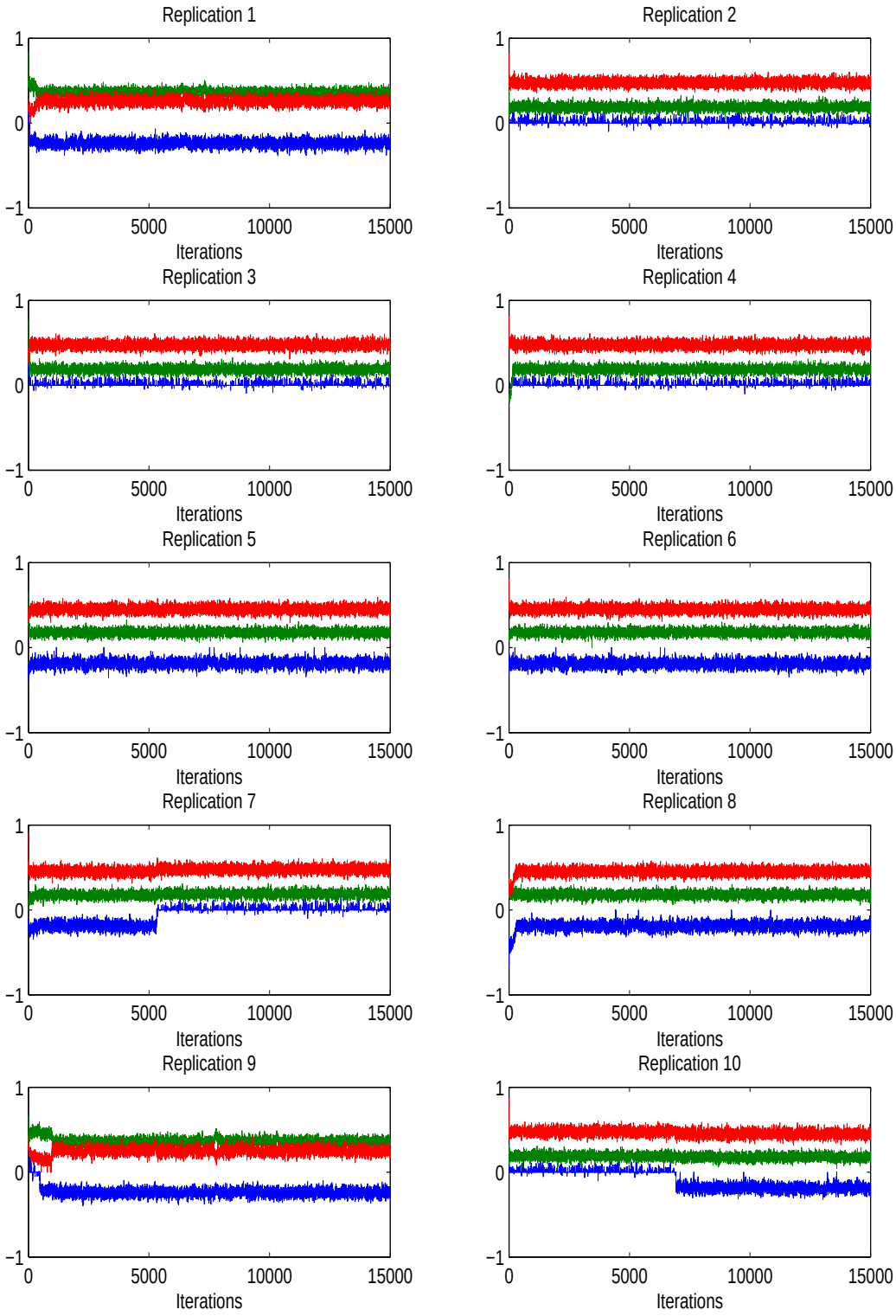

Figure 4.7: First 15,000 iterations from 10 sequences of the factor loadings of one randomly selected variable from the sampler of Kaufmann and Schumacher (2013) for simulated data with approximately sparse structure in Λ, using different starting points.

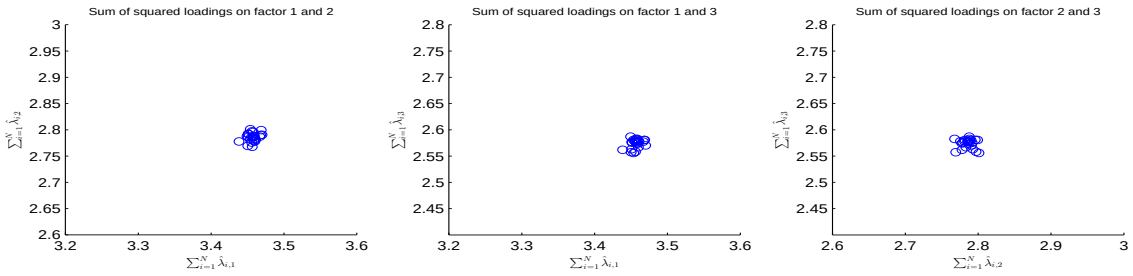

Figure 4.8: Sum of squared loadings per factor calculated from the posterior estimate $\hat{\Lambda}$ for simulated data with 73% zero elements in Λ, using different starting points.

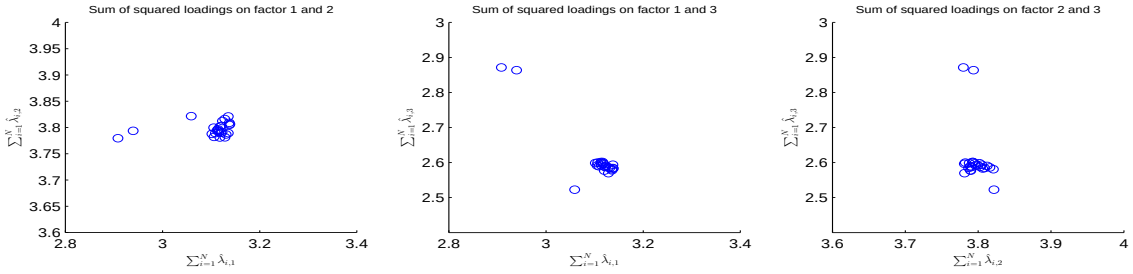

Figure 4.9: Sum of squared loadings per factor calculated from the posterior estimate $\hat{\Lambda}$ for simulated data with 55% zero elements in Λ, using different starting points.

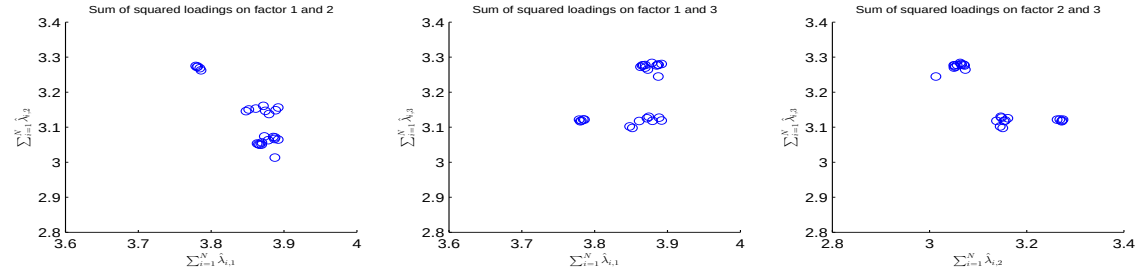

Figure 4.10: Sum of squared loadings per factor calculated from the posterior estimate $\hat{\Lambda}$ for simulated data with 34% zero elements in Λ, using different starting points.

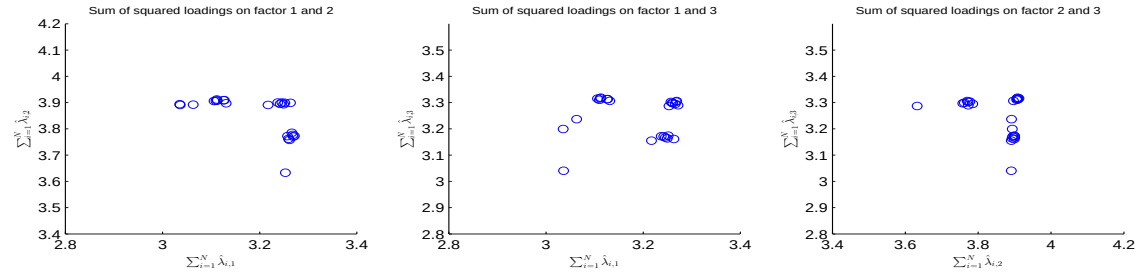

Figure 4.11: Sum of squared loadings per factor calculated from the posterior estimate $\hat{\Lambda}$ for simulated data with approximately sparse structure in Λ, using different starting points.

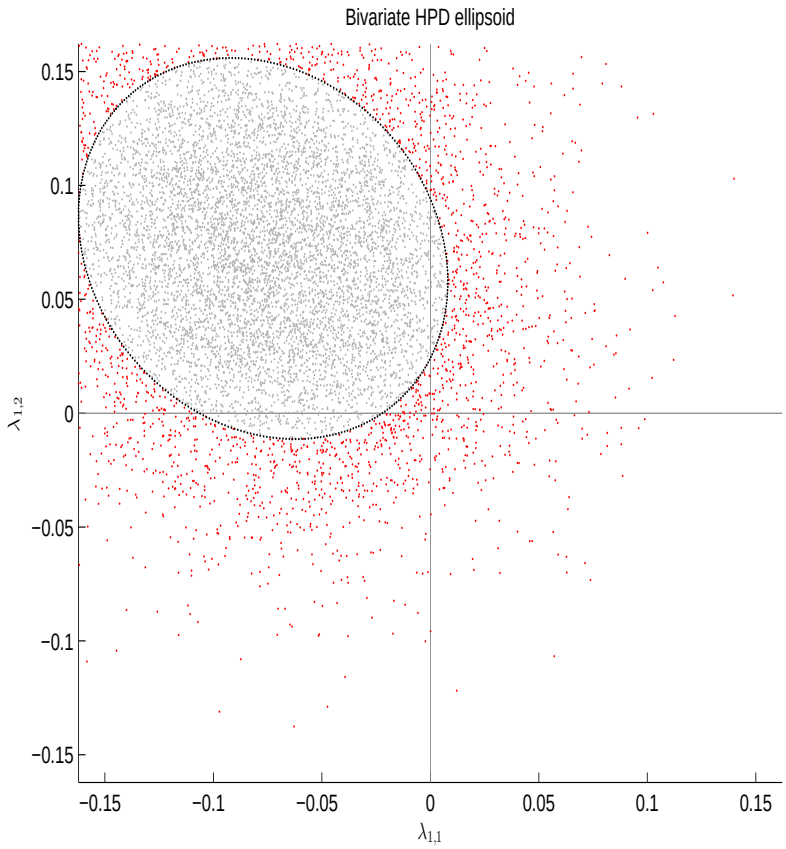

Figure 4.12: $1 - \alpha$ HPD ellipsoid in two dimensions.

Notes: Grey dots are within the HPDE, red dots are outside the HPDE. The boundary line is shown for illustration purposes only.

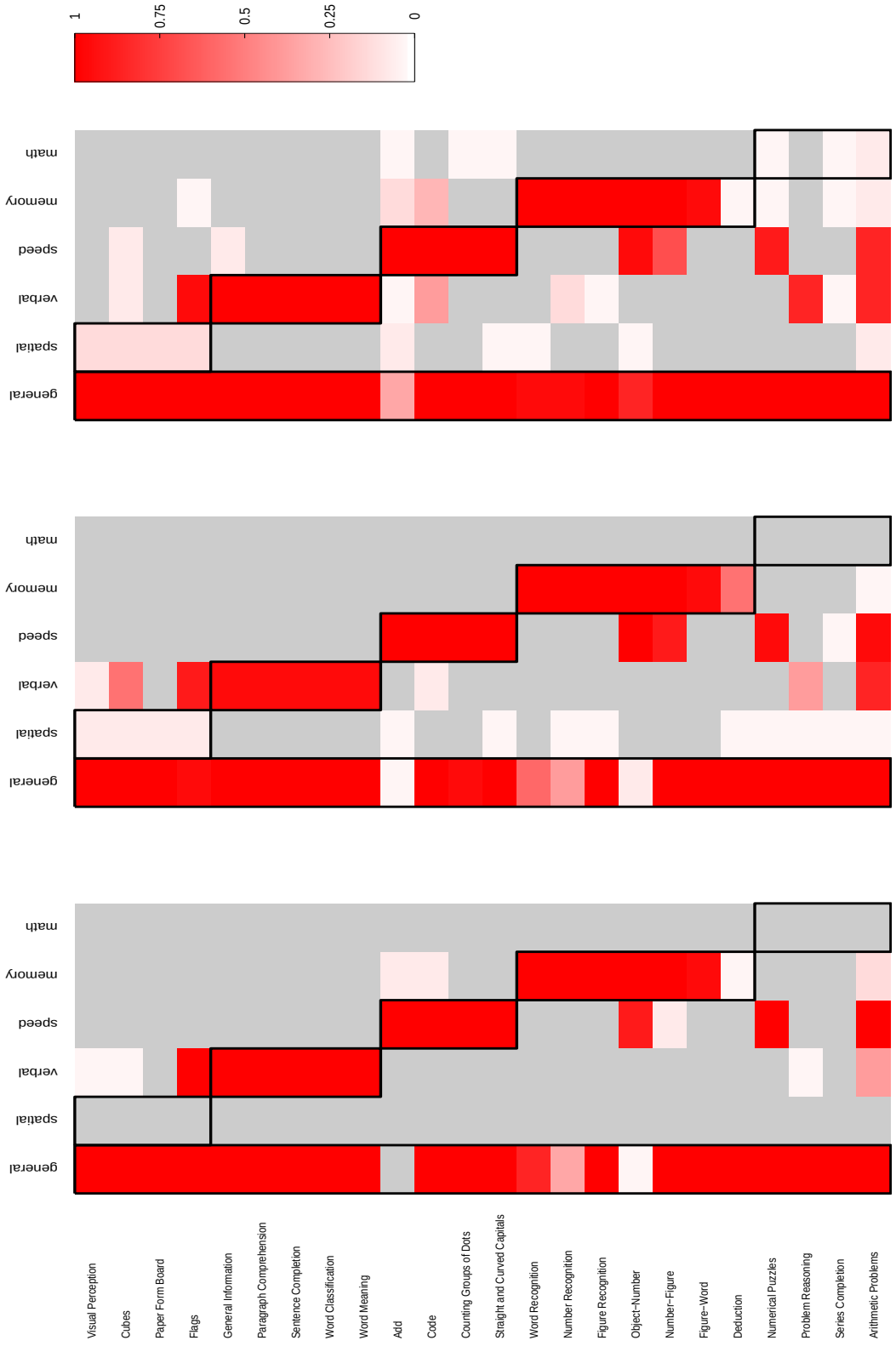

Figure 4.13: Mean association probabilities, calculated as averages over 20 model estimates from the two-step approach for $\alpha = 0.01$ (left), $\alpha = 0.05$ (middle) and $\alpha = 0.1$ (right).

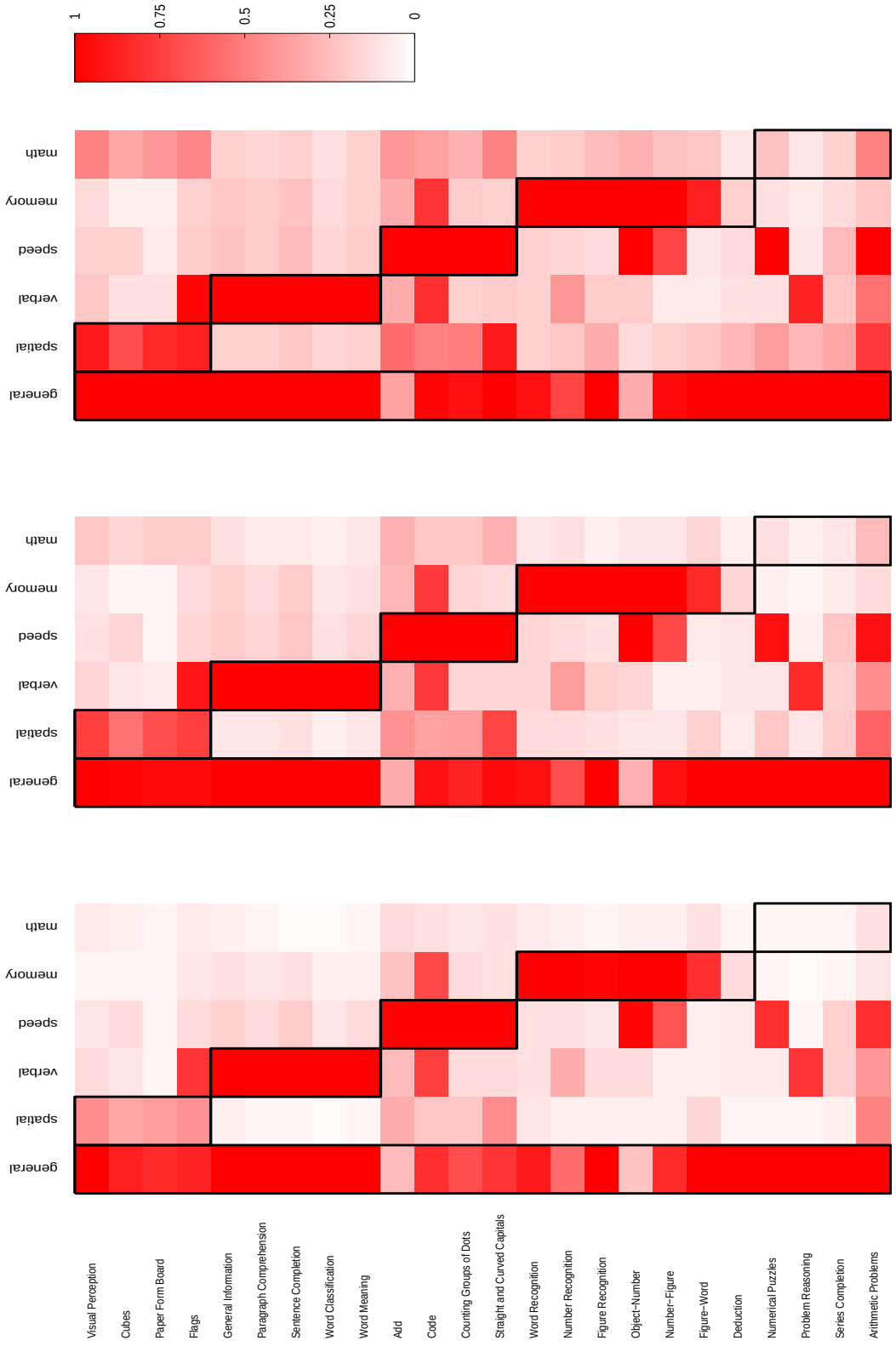

Figure 4.14: 16% (left), 50% (middle) and 84% quantiles (right) of the 25 estimates for the association probabilities from the sparse sampler.

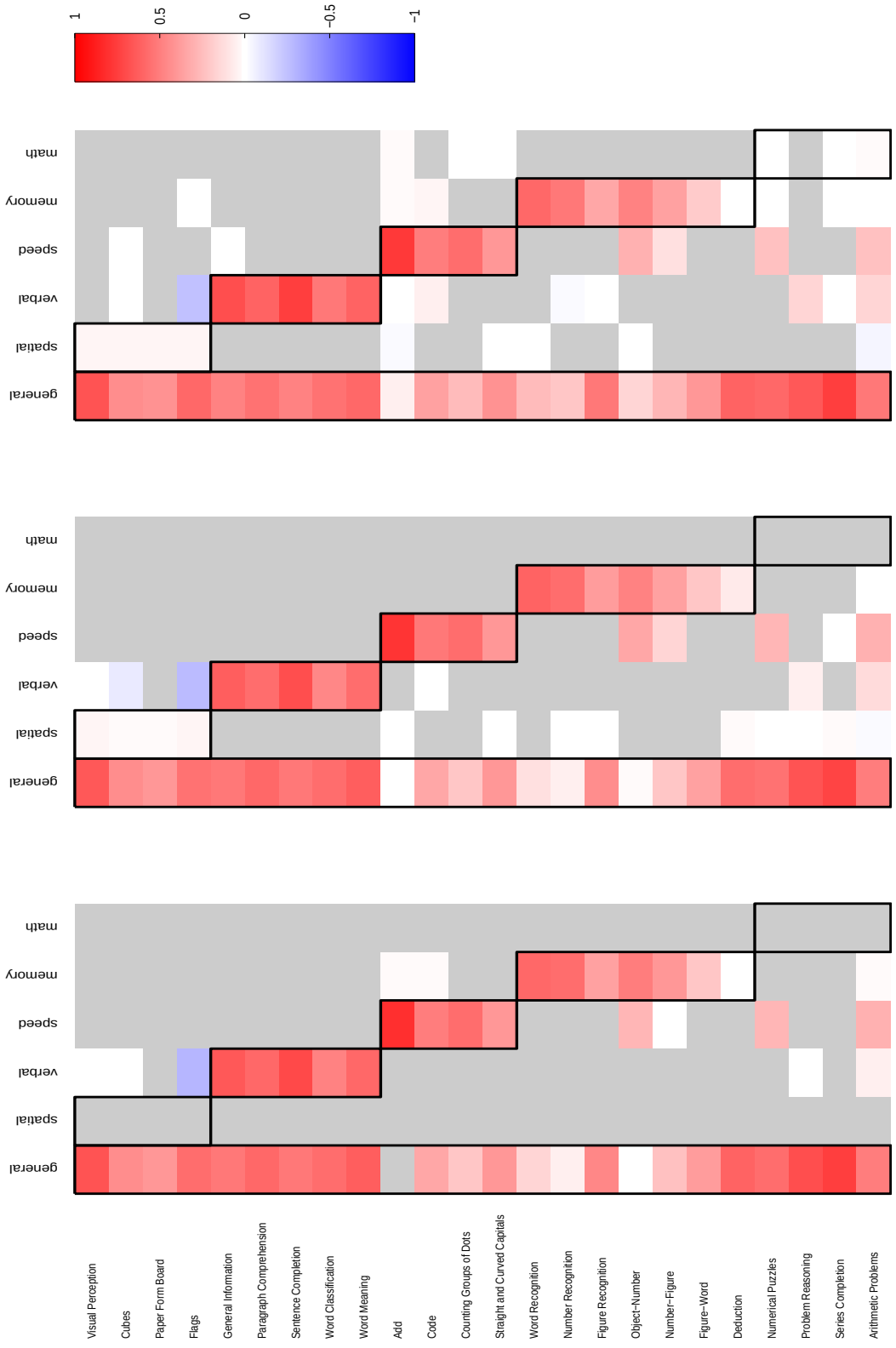

Figure 4.15: Loadings in the parsimonious structure obtained from the two-step approach.

Notes: The figure shows the means over 20 model estimates for $\alpha = 0.01$ (left), $\alpha = 0.05$ (middle) and $\alpha = 0.1$ (right). Elements identified as zero are shown as grey patches.

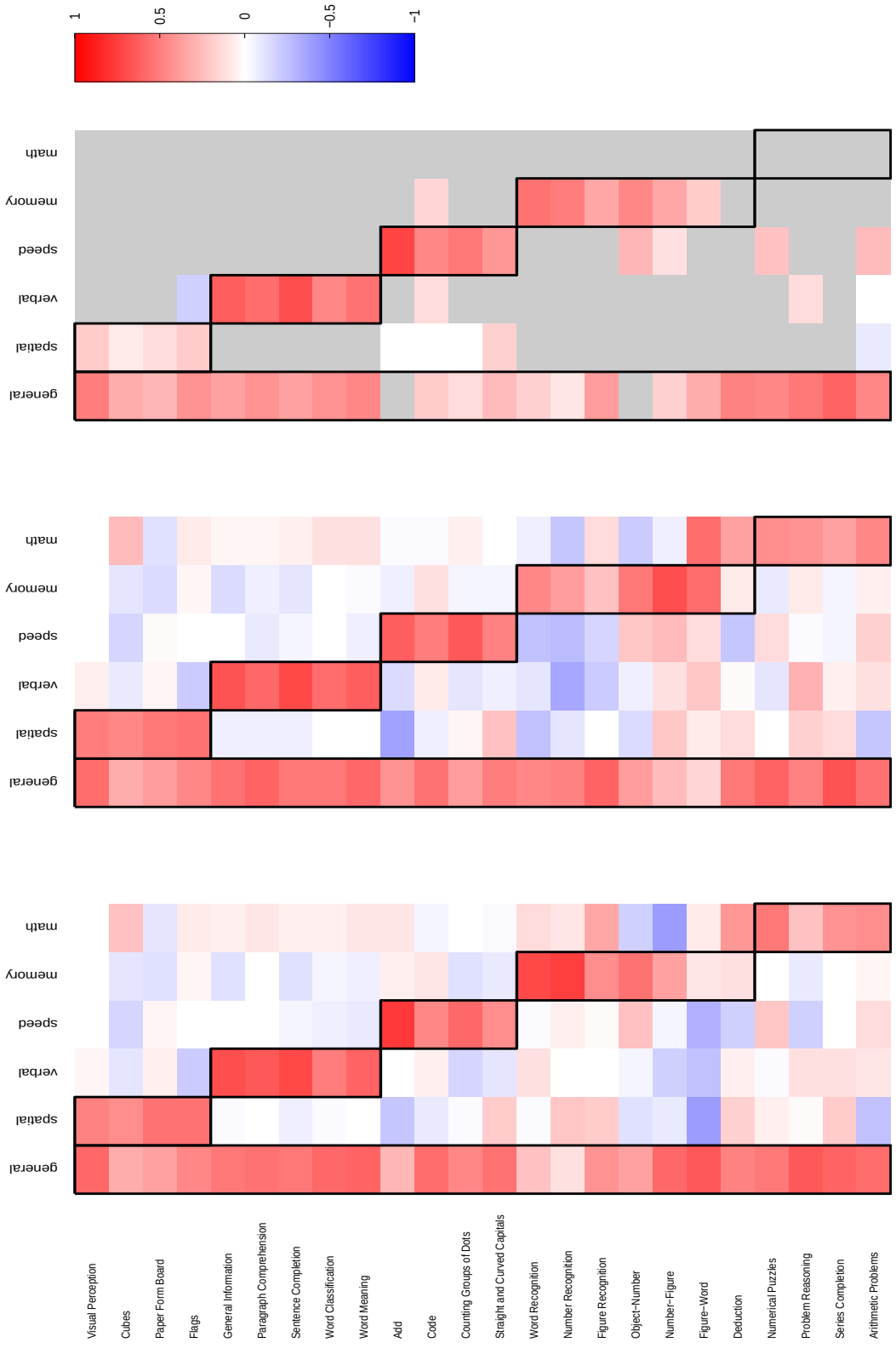

Figure 4.16: Loadings in the parsimonious structure obtained from the benchmarks.

Notes: PC factor analysis rotated to the assumed bi-factor structure (left), WOP approach rotated to the assumed bi-factor structure (middle) and sparse sampler (right). Elements identified as zero are shown as grey patches.

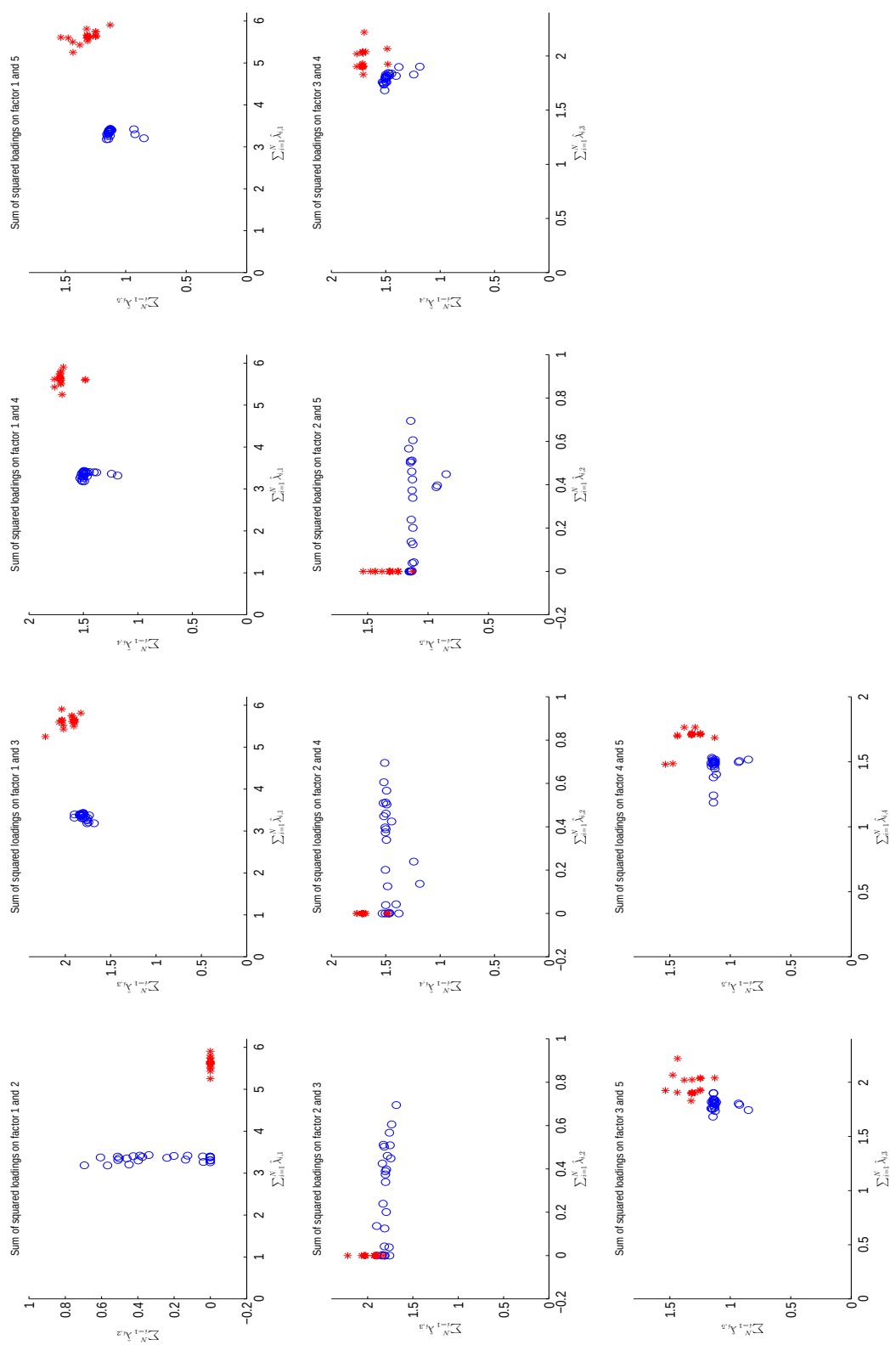

Figure 4.17: Sum of squared loadings per factor calculated from the posterior estimates for $\hat{\Lambda}$ for the results from the sparse sampler (blue) and for the results from the two-step approach with $\alpha = 0.01$ (red).

Appendix 4.A : The Unconstrained Gibbs Sampler

The Gibbs sampler for the static factor model largely follows the setup of Otrok and Whiteman (1998) and Kose et al. (2003), but omits the parameters governing the dynamics in the factors and the according filtering, or quasi-differencing, steps that are part of the dynamic single- and multi-factor models discussed there. The chosen prior distributions are conjugate and independent, where the prior distribution of the loadings is a matrix normal distribution that is itself the product of independent K-variate normal distributions, and the prior distribution of the idiosyncratic error covariances is an inverse Wishart distribution that is itself the product of independent univariate inverse gamma distributions, hence

$$\pi(\Lambda, \Sigma) = \pi(\Lambda)\pi(\Sigma) = \prod_{i=1}^{N} f_N(\lambda_i|\mu_{\lambda_i}, \Sigma_{\lambda_i}) \prod_{i=1}^{N} f_{IG}(\sigma_i^2|\underline{\alpha}_i, \underline{\beta}_i), \tag{4.49}$$

with

$$\pi(\Lambda) = \prod_{i=1}^{N} (2\pi)^{-\frac{K}{2}} |\Sigma_{\lambda_i}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}((\lambda_i - \mu_{\lambda_i})'\Sigma_{\lambda_i}^{-1}(\lambda_i - \mu_{\lambda_i}))\right\}, \tag{4.50}$$

and

$$\pi(\Sigma) = \prod_{i=1}^{N} \frac{\underline{\beta}_i^{\underline{\alpha}_i}}{\Gamma(\underline{\alpha}_i)} \sigma_i^{-2(\underline{\alpha}_i+1)} \exp\left\{-\frac{\underline{\beta}_i}{\sigma_i^2}\right\}. \tag{4.51}$$

To obtain a sample from an orthogonally mixing posterior distribution that is fit to be postprocessed with the WOP algorithm explained in Appendix 4.C, the chosen prior distributions have to be invariant under orthogonal transformations. To guarantee orthogonal invariance of the prior, μ_{λ_i} must be a $K \times 1$ vector of zeros, and Σ_{λ_i} must be a multiple of the identity matrix, hence $\Sigma_{\lambda_i} = \underline{c}_i I_K$. With the likelihood invariant under orthogonal transformations of Λ and $\{f_t\}_{t=1}^T$ and the prior distributions likewise orthogonally invariant, the posterior distributions are also orthogonally invariant.

In order to enable the Gibbs sampler to generate a sample from the posterior density of the model parameters of interest Λ and Σ, the latent factors are likewise sampled and are used in a data augmentation step, see Tanner and Wong (1987). The prior distribution of the factors is a K-variate normal distribution,

$$\pi\left(\{f_t\}_{t=1}^T\right) = \prod_{t=1}^{T} f_N(f_t|\mu_f, \Sigma_f), \tag{4.52}$$

with μ_f is a $K \times 1$ vector of zeros, implying that the factors have zero mean, and $\Sigma_f = I_K$ to fix the scaling of the factors and loadings.

The Gibbs sampler for the static factor model proceeds as follows for every iteration $z \in \{1, \ldots, Z\}$:

1. Sample $\Lambda^{(z)}$ from its full conditional distribution $\lambda_i^{(z)}|\{f_t^{(z-1)}\}_{t=1}^T, \Sigma^{(z-1)}; Y$ for all $i \in \{1, \ldots, N\}$.
2. Sample $\Sigma^{(z)}$ from its full conditional distribution $\Sigma^{(z)}|\{f_t^{(z-1)}\}_{t=1}^T, \Lambda^{(z)}; Y$.
3. Sample the factors from their full conditional distribution $\{f_t^{(z)}\}_{t=1}^T|\Lambda^{(z)}, \Sigma^{(z)}; Y$.

The full conditional distribution of the loadings is

$$g(\Lambda|\{f_t\}_{t=1}^T, \Sigma, \{y_t\}_{t=1}^T) = \prod_{i=1}^N (2\pi)^{-\frac{K}{2}} |\Sigma_{\lambda_i}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\lambda_i - \mu_{l_i})' \Sigma_{l_i} (\lambda_{\lambda_i} - \mu_{l_i})\right), \quad (4.53)$$

where $\Sigma_{l_i} = \left(\sigma_i^{-2} \sum_{t=1}^T f_t f_t' + \Sigma_{\lambda_i}^{-1}\right)^{-1}$ and $\mu_{l_i} = \Sigma_{l_i}\left(\Sigma_{\lambda_i}^{-1}\mu_{\lambda_i} + \sigma_i^{-2}\sum_{t=1}^T f_t y_{it}\right)$, the full conditional distribution of the idiosyncratic variances is

$$g\left(\Sigma|\{f_t\}_{t=1}^T, \Lambda, \{y_t\}_{t=1}^T\right) = \prod_{i=1}^N \frac{b_i^{a_i}}{\Gamma(a_i)} (\sigma_i)^{-2a_i-1} \exp\left(-\frac{b_i}{\sigma_i^2}\right), \quad (4.54)$$

where $a_i = \frac{T}{2} + \underline{\alpha}_i$ and $b_i = \frac{1}{2}\sum_{t=1}^T (y_t - \lambda_i' f_t)^2 + \underline{\beta}_i$, and the full conditional distribution of the factors is

$$g\left(\{f_t\}_{t=1}^T|\Lambda, \Sigma, \{y_t\}_{t=1}^T\right) = \prod_{t=1}^T (2\pi)^{-\frac{K}{2}} |\Sigma_{f_t}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(f_t - \mu_{f_t})' \Sigma_{f_t}^{-1} (f_t - \mu_{f_t})\right), \quad (4.55)$$

where $\Sigma_{f_t} = (\Lambda'\Sigma^{-1}\Lambda + \Sigma_f^{-1})^{-1} = (\Lambda'\Sigma^{-1}\Lambda + I_K)^{-1}$ and $\mu_{f_t} = \Sigma_{f_t}(\Sigma_f^{-1}\mu_f + \Lambda'\Sigma^{-1}y_t) = \Sigma_{f_t}(\Lambda'\Sigma^{-1}y_t)$.

The prior hyperparameters to be chosen are therefore $\{\underline{\alpha}_i, \underline{\beta}_i, \underline{c}_i\}_{i=1}^N$.

Appendix 4.B: Orthogonal Mixing in the Gibbs Sampler

When no constraints are imposed to solve the identification problem, the sampler described in Appendix 4.A is orthogonally mixing. This implies that in each iteration of the sampler, the sample space of Λ may undergo an orthogonal transformation, which can be expressed by the matrix D. In the following, the effect of such a transformation on the moments of the full conditional distributions given in Appendix 4.A is shown.

The parameters of the full conditional distribution of Λ in Equation (4.53) become

$$\begin{aligned}\Sigma_{l_i}|D &= \left(\sigma_i^{-2}\sum_{t=1}^T D' f_t (D' f_t)' + \underline{c}_i^{-2} I_K\right)^{-1} \\ &= D'\left(\left(\sigma_i^{-2}\sum_{t=1}^T f_t f_t' + \underline{c}_i^{-2} I_K\right)^{-1}\right) D = D'\Sigma_{l_i} D \end{aligned} \quad (4.56)$$

and

$$\begin{aligned}\mu_{l_i}|D &= D'\Sigma_{l_i}D\left(\sigma_i^{-2}\sum_{t=1}^{T}(D'f_t)y_{it}\right)\\ &= D'\Sigma_{l_i}(\Lambda\Sigma^{-1}y_t) = D'\mu_{l_i}.\end{aligned} \tag{4.57}$$

The parameters of the full conditional distribution of the factors in Equation (4.55) accordingly become

$$\begin{aligned}\Sigma_{f_t}|D &= ((\Lambda D)'\Sigma^{-1}(\Lambda D) + I_K)^{-1}\\ &= D'(\Lambda'\Sigma^{-1}\Lambda + I_K)^{-1}D = D'\Sigma_{f_t}D\end{aligned} \tag{4.58}$$

and

$$\begin{aligned}\mu_{f_t}|D &= D'\Sigma_{f_t}D((\Lambda D)'\Sigma^{-1}y_t)\\ &= D'\Sigma_{f_t}(\Lambda\Sigma^{-1}y_t) = D'\mu_{f_t}.\end{aligned} \tag{4.59}$$

The parameters of the full conditional distribution of the idiosyncratic variances in Equation (4.54) remain unchanged, as the effect of an orthogonal transformation of Λ cancels out against that of an according orthogonal transformation of $\{f_t\}_{t=1}^T$.

Appendix 4.C: The Weighted Orthogonal Procrustes Algorithm

The weighted orthogonal Procrustes algorithm solves the following optimization problem for a sequence of draws from an orthogonally mixing posterior distribution of Λ, which is denoted as $\{\Lambda^{(z)}\}_{z=1}^Z$:

$$\{\{D^{(z)}\}_{z=1}^Z, \Lambda^*\} = \arg\min \sum_{z=1}^{Z} L_D(\Lambda^*, \Lambda^{(z)}(D^{(z)})) \quad \text{s.t.} \quad D^{(z)\prime}D^{(z)} = I,\ z = 1, \ldots, Z, \tag{4.60}$$

where L_D denotes the quadratic loss function

$$L_D(\Lambda^*, \Lambda^{(z)}(D^{(z)})) = \operatorname{tr}\left[(\Lambda^{(z)}(D^{(z)}) - \Lambda^*)'(\Lambda^{(z)}(D^{(z)}) - \Lambda^*)\right] \tag{4.61}$$

$$= \operatorname{tr}\left[(\Lambda^{(z)}D^{(z)} - \Lambda^*)'(\Lambda^{(z)}D^{(z)} - \Lambda^*)\right]. \tag{4.62}$$

$$\tag{4.63}$$

A solution to this optimization problem is obtained iteratively by the WOP algorithm. The algorithm iterates between the following two steps until convergence, i.e. until the change in Λ^* between two subsequent iterations, measured as the matrix norm of the difference, falls below

a defined threshold value, say 10^{-9}. At the beginning, an initialization for Λ^* is required, where $\Lambda^{(Z)}$ is chosen for convenience. As Step 1 minimizes the (weighted) distance between the transformed draws and the given Λ^*, it provides an unique orientation to each sampled $\Lambda^{(z)}$. In Step 2, the estimator is determined based on an orientated sample. For arbitrary initial choices of Λ^* taken from the unconstrained sampler output, less than ten iterations usually suffice to achieve convergence to a fixed point Λ^*.

Step 1 For given Λ^* the following minimization problem for $D^{(z)}$ has to be solved for each $z = 1, \ldots, Z$:

$$D^{(z)} = \arg\min L_{D,1}(\Theta^*, \Theta^{(z)}(D^{(z)})) \quad \text{s.t.} \quad D^{(z)\prime} D^{(z)} = I. \tag{4.64}$$

The solution of this orthogonal Procrustes (OP) problem is provided by Kristof (1964) and Schönemann (1966), see also Golub and van Loan (2013). It involves the following calculations:

1.1 Define $S_z = \Lambda^{(z)\prime}\Lambda^*$.

1.2 Do the singular value decomposition $S_z = U_z M_z V_z'$, where U_z and V_z denote the matrix of eigenvectors of $S_z S_z'$ and $S_z' S_z$, respectively, and M_z denotes a diagonal matrix of singular values, which are the square roots of the eigenvalues of $S_z S_z'$ and $S_z' S_z$. Note that the eigenvalues of $S_z S_z'$ and $S_z' S_z$ are identical.

1.3 Obtain the orthogonal transformation matrix $D^{(z)} = U_z V_z'$.

For further details on the derivation of this solution, see Schönemann (1966) or Appendix 2.A.

Note that if the dispersion between the cross sections is rather large, the solution may be improved by considering weights, turning the problem to be solved into a weighted orthogonal Procrustes (WOP) problem, see e.g. Lissitz et al. (1976) and Koschat and Swayne (1991). Thus Step 1.1 above is altered into

1.1a Define $S_z = \Lambda^{(z)\prime} W \Lambda^*$,

where the weighting matrix W has to be diagonal with strictly positive diagonal elements and is initialized as the inverses of the estimated lengths of the loading vectors, i.e.

$$W = Z \left(\sum_{z=1}^{Z} \sqrt{(\Lambda^{(z)} \Lambda^{(z)\prime}) \odot I_N} \right)^{-1}. \tag{4.65}$$

The weighting function used here is a function of the number of factors and the determinants of the estimated covariance matrices, which are a measure invariant to orthogonal transformations, i.e. $W = \text{diag}(w_1, \ldots, w_N)$, where

$$w_i = \det \left(\frac{1}{Z} \sum_{z=1}^{Z} (\lambda_i^{(z)} - \lambda_i^*)(\lambda_i^{(z)} - \lambda_i^*)' \right)^{-\frac{1}{K}}, \quad i = 1, \ldots, N. \tag{4.66}$$

The weighting scheme scales the loadings in such a way that the estimated covariance matrix has determinant 1 for each variable.

Step 2 Choose Λ^* as

$$\Lambda^* = \frac{1}{Z}\sum_{z=1}^{Z}\Lambda^{(z)}D^{(z)}. \tag{4.67}$$

Chapter 5

Application to the German Labor Market

5.1 Introduction

Investigation of the common movements of business cycles across different countries, or different regions within the same country, has found much research interest over the past years. The major questions revolve around the identification and estimation of common business cycles, whether these cycles are converging or decoupling, and what lies behind the regional differences. Factor models as well as clustering approaches have recently played an important role in answering these questions. Kose et al. (2003) use a structural factor model to decompose the business cycle for various countries into global, regional and country-specific components. Eickmeier and Breitung (2006) likewise apply a structural factor model to investigate whether the central and eastern European economies that joined the European Union (EU) in 2004 qualify for joining the European Monetary Union (EMU) based on the synchronization of their business cycles with those of the EMU countries. Using a Markov switching approach, Owyang et al. (2005) analyze the behavior of individual states of the United States (U.S.) and find differences in their exposure to the nationwide business cycle as well as differences in the business cycle timing. Hamilton and Owyang (2012) use a clustering approach to determine which states in the U.S. share a common pattern with respect to the characteristics of their business cycles. A principal-components based factor model approach alternatively combined with spatial lags or spatial errors is proposed by Artis et al. (2011), who find that the relation between nationwide and regional business cycle in the U.S. and the EU is similar and stable over time. Using a structural factor model, Kose et al. (2008) find convergence both within the group of industrialized countries and within the group of emerging market economies, but divergence between the two groups. In a recent extension to the approach proposed by Kose et al. (2003), Francis et al. (2012) propose a combination of a clustering approach and a structural factor model approach where the cluster membership is determined within the model instead of imposing it ex ante.

In the present chapter, I attempt to identify a common business cycle for Germany, as well as regional-specific variations. To this end, I estimate a latent dynamic factor model, using the weighted orthogonal Procrustes (WOP) approach from Chapter 3. The model is allowed to include dynamic error components, hence following the specification of Bai and Ng (2007), so

the WOP estimation approach from Chapter 3 is accordingly extended. Afterwards, I apply the two-step approach to find a parsimonious loadings structure from Chapter 4 to determine whether some of the latent factors are constrained to certain regions only.

As regional GDP data are usually not available, particularly at monthly frequency, researchers have resorted to other variables, most notably employment figures. Conversely, policymakers may consider employment as the variable of interest rather than regional GDP. The seminal paper of Blanchard and Katz (1992) hence focuses directly on labor market dynamics, Korobilis and Gilmartin (2011) measure the effect of monetary policy on state-level unemployment, and Owyang et al. (2013) analyze the synchronization of employment expansion and contraction phases for 58 large U.S. cities by means of a Markov switching model. Hamilton and Owyang (2012) likewise use employment figures at quarterly frequency as an instrument to measure regional business cycles.[1] In order to find out about the extent of regional co-movements of the business cycle, I therefore use absolute unemployment figures, which are available at monthly frequency for Germany's 402 counties, or NUTS-3 regions. Based on these figures, growth rates of unemployment are calculated, as in Boysen-Hogrefe and Pape (2011), which are then used in the analysis.

The remainder of this chapter is structured as follows. Section 5.2 discusses the model setup and the question of model identification. Section 5.3 describes the sampling approach and provides two model selection criteria. Section 5.4 describes the data set in more detail, Section 5.5 discusses the results of the factor analysis with the full loadings matrix, while Section 5.6 accordingly discusses the results for the sparse model. Section 5.7 presents the results of a brief forecasting exercise comparing the full and sparse models, and Section 5.8 concludes.

5.2 Model Setup and Identification

Consider a dynamic factor model with K factors, written as

$$y_t = \sum_{s=0}^{S} \Lambda_s f_{t-s} + e_t \quad \text{for } t \in \{1, \dots, T\}, \tag{5.1}$$

as e.g. in Bai and Ng (2007), where y_t is an $N \times 1$ vector of demeaned observed data for $t \in \{1, \dots, T\}$, Λ_s for $s \in \{0, \dots, S\}$ are $N \times K$ matrices of loadings on the contemporaneous and lagged latent factors, f_{t-s} for $s \in \{0, \dots, S\}$ are the $K \times 1$ vectors of contemporaneous and lagged latent factors, which follow a vector-autoregressive (VAR) process of order P, so

$$f_t = \sum_{p=1}^{P} \Phi_p f_{t-p} + \epsilon_t \quad \text{for } t \in \{1, \dots, T\}. \tag{5.2}$$

[1] For Germany, GDP data on NUTS-3 level (nomenclature des unités territoriales statistiques) is in fact available at annual frequency. This data is not suitable for the factor model approach, however, due to the very small time dimension. Juessen (2009) uses this data for a nonparametric dynamic kernel estimation approach.

The idiosyncratic error terms follow a VAR process of order Q, so

$$e_t = \sum_{q=1}^{Q} \Theta_q e_{t-q} + \xi_t \quad \text{for } t \in \{1, \ldots, T\}, \tag{5.3}$$

where throughout, all Θ_q are assumed to be diagonal, implying independent autoregressive (AR) processes, hence the idiosyncratic errors for all series are mutually independent, also known as the Frisch case, after Frisch (1934), see Scherrer and Deistler (1998).[2] Denoting the i^{th} diagonal element of Θ_q as $\theta_{q,i,i}$, the idiosyncratic error terms for each series can therefore be written as

$$e_{i,t} = \sum_{q=1}^{Q} \theta_{q,i,i} e_{t-q} + \xi_{i,t} \quad \text{for } t \in \{1, \ldots, T\}. \tag{5.4}$$

5.2.1 Model Identification up to the Rotation Problem

Note that for finite S, this dynamic factor model can also be written as a static factor model, hence

$$y_t = \Lambda F_t + e_t \quad \text{for } t \in \{1, \ldots, T\}, \tag{5.5}$$

where $\Lambda = [\Lambda_0, \ \ldots, \ \Lambda_S]$ and $F_t = [f_t', \ \ldots, \ f_{t-S}']'$. Regarding the innovations in the factors and idiosyncratic components, it is assumed that

$$\begin{pmatrix} \epsilon_t \\ \xi_t \end{pmatrix} \sim i.i.d. \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & 0 \\ 0 & \Sigma \end{pmatrix} \right). \tag{5.6}$$

Thus, the full model can also be written as

$$\begin{aligned} y_t &= \sum_{s=0}^{S} \left(\Lambda_s \left(\sum_{p=1}^{P} \Phi_p f_{t-s-p} + \epsilon_{t-s} \right) \right) + \sum_{q=1}^{Q} \Theta_q e_{t-q} + \xi_t \\ &= \sum_{s=0}^{S} \left(\Lambda_s \left(I_K - \sum_{p=1}^{P} \Phi_p L^p \right)^{-1} \epsilon_{t-s} \right) + \left(I_N - \sum_{q=1}^{Q} \Theta_q L^q \right)^{-1} \xi_t \\ &= \sum_{s=0}^{S} \left(\Lambda_s (\Phi(L))^{-1} \epsilon_{t-s} \right) + (\Theta(L))^{-1} \xi_t., \end{aligned} \tag{5.7}$$

where L is the lag operator and

$$\Phi(L) = -\sum_{p=0}^{P} \Phi_p L^p, \quad \text{and} \quad \Theta(L) = -\sum_{q=0}^{Q} \Theta_q L^q, \tag{5.8}$$

[2] In fact, the Frisch case implies that all off-diagonal elements of $\Theta(L)$ in Equation (5.8) are zero, which is also possible for conformably chosen nonzero off-diagonal elements for the Θ_q.

where $\Phi_0 = -I_K$ and $\Theta_0 = -I_N$. The inverses of $\Theta(L)$ and $\Phi(L)$ are then lag polynomials of moving-average processes of infinite order in ϵ_{t-s} and ξ_t, respectively. By $\Phi_0 = -I_K$ and the assumption that

$$E[\epsilon_t \epsilon_t'] = I_K, \tag{5.9}$$

the scaling of the innovations in the factors is fixed, while

$$E[D\epsilon_t \epsilon_t' D'] = I_K \tag{5.10}$$

holds for every orthogonal matrix $D \in O(K)$. Thus consider an orthogonal transformation by a matrix D of the factors and loadings matrices in Equation (5.1), which is

$$y_t = \sum_{s=0}^{S} \Lambda_s D D' f_{t-s} + e_t \quad \text{for } t \in \{1, \dots, T\}. \tag{5.11}$$

Accordingly, consider an orthogonal transformation of the stacked static factor model representation in Equation (5.5), which is

$$y_t = \Lambda(I_{S+1} \otimes D)(I_{S+1} \otimes D')F_t + e_t \quad \text{for } t \in \{1, \dots, T\}. \tag{5.12}$$

The orthogonally transformed factors in Equation (5.11) then evolve as

$$D' f_t = \sum_{p=1}^{P} D'\Phi_p D D' f_{t-p} + D'\epsilon_t \quad \text{for } t \in \{1, \dots, T\}. \tag{5.13}$$

The equivalent of Equation (5.7) for the orthogonally transformed factors and loadings is then

$$\begin{aligned} y_t &= \sum_{s=0}^{S} \left(\Lambda_s D \left(\sum_{p=1}^{P} D'\Phi_p D D' f_{t-s-p} + D'\epsilon_{t-s} \right) \right) + \sum_{q=1}^{Q} \Theta_q e_{t-q} + \xi_t \\ &= \sum_{s=0}^{S} \left(\Lambda_s D \left(I_K - \sum_{p=1}^{P} D'\Phi_p D L^p \right)^{-1} D'\epsilon_{t-s} \right) + \left(I_N - \sum_{q=1}^{Q} \Theta_q L^q \right)^{-1} \xi_t \\ &= \sum_{s=0}^{S} \left(\Lambda_s D (\Phi(L, D))^{-1} \epsilon_{t-s} \right) + (\Theta(L))^{-1} \xi_t., \end{aligned} \tag{5.14}$$

where

$$\Phi(L, D) = -\sum_{p=0}^{P} D'\Phi_p D L^p. \tag{5.15}$$

Next, denote all matrices $\tilde{\Lambda}$ where

$$\tilde{\Lambda} = \Lambda(I_{S+1} \otimes D) \quad \text{with} \quad D \in O(K), \tag{5.16}$$

i.e. D is an arbitrary orthogonal matrix, as the equivalence class of Λ, and all lag polynomials $\tilde{\Phi}(L)$ where

$$\tilde{\Phi}(L) = \Phi(L, D) \quad \text{with} \quad D \in O(K), \tag{5.17}$$

as the equivalence class of $\Phi(L)$. Then for the identification problem to be reduced to the rotation problem, all $(\tilde{\Phi}(L))^{-1}$ must be left coprime, i.e. its greatest left divisor must be I_K, see e.g. Hannan and Deistler (2012), Chapter 2. If some $(\tilde{\Phi}(L))^{-1}$ exists that is not left coprime, a matrix U must exist, such that $(\tilde{\Phi}(L))^{-1}$ can be replaced by $U^{-1}(\tilde{\Phi}(L))^{-1}$, and conversely, Λ can be replaced by $\Lambda(I_{S+1} \otimes U)$.

5.2.2 The State-Space Representation

When looking at model identification, it is also useful to look at the state-space representation of the factor model. To account for the serial correlation in the errors, the data is accordingly filtered, as e.g. in Otrok and Whiteman (1998). To filter the first Q observations per variable, the companion matrix for the AR process in the idiosyncratic error terms for variable i is constructed as

$$M_i = \begin{pmatrix} \theta_{i,i,1} & \dots & \theta_{i,i,Q-1} & \theta_{i,i,Q} \\ 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix}, \tag{5.18}$$

see e.g. Hamilton (1994), Chapter 10. The stationary covariance matrix of the errors in the first Q observations per variable in vectorized form is then $\Sigma_{Q_i}\sigma_i^2$ with

$$\text{vec}(\Sigma_{Q_i}) = (I_{Q^2} - M_i \otimes M_i)^{-1}\text{vec}(u_1'u_1) \quad \text{for} \;\; i \in \{1, \dots, N\}, \tag{5.19}$$

where $u_1 = [1 \;\; 0 \;\; \dots 0]'$ is the first Q^2-dimensional canonical unit vector. With

$$\Sigma_{Q_i} = C_i'C_i \tag{5.20}$$

being the Cholesky decomposition of Σ_{Q_i}, the filtered first Q observations for variable i obtain as

$$[y_{i,1}^* \;\; \dots \;\; y_{i,Q}^*]' = C_i^{-1}[y_{i,1} \;\; \dots \;\; y_{i,Q}]' \quad \text{for} \;\; i \in \{1, \dots, N\}, \tag{5.21}$$

and the filtered remaining $T - Q$ observations for all N variables obtain as

$$y_t^* = y_t - \sum_{q=1}^{Q} \Theta_q y_{t-q} \quad \text{for } t \in \{Q+1, \dots, T\}. \tag{5.22}$$

Next, let $R = \max\{P, Q+1\}$. The observation equation is then

$$y_t^* = \underbrace{[\Lambda \quad -\Theta_1\Lambda \quad \dots \quad -\Theta_Q\Lambda \quad \mathbf{0}_\Lambda]}_{\mathbf{\Lambda}} \underbrace{[F_t' \quad \dots \quad F_{t-R+1}']'}_{\mathbf{F_t}} + \xi_t \quad \text{for } t \in \{1, \dots, T\}, \tag{5.23}$$

where $\mathbf{\Lambda}$ is an $N \times K(S+1)R$ matrix and $\mathbf{F_t}$ is a $K(S+1)R \times 1$ vector. The matrix $\mathbf{0}_\Lambda = 0_{N \times K(S+1)(R-Q-1)}$ is a matrix of zeros of dimension $N \times K(S+1)(R-Q-1)$. The state equation for $\mathbf{F_t}$ is accordingly

$$\begin{aligned} \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-R+1} \end{pmatrix} &= \begin{pmatrix} \mathbf{\Phi_1} & \dots & \dots & \mathbf{\Phi_R} \\ I_{K(S+1)} & & \mathbf{0}_\Phi & \mathbf{0}_\Phi \\ & \ddots & & \vdots \\ \mathbf{0}_\Phi & & I_{K(S+1)} & \mathbf{0}_\Phi \end{pmatrix} \begin{pmatrix} F_{t-1} \\ F_{t-2} \\ \vdots \\ F_{t-R} \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ 0_\epsilon \end{pmatrix} \\ &= \mathbf{F_t} = \mathbf{H}\mathbf{F_{t-1}} + \epsilon_\mathbf{t} \quad \text{for } t \in \{1, \dots, T\}, \end{aligned} \tag{5.24}$$

where $\mathbf{0}_\Phi = 0_{K(S+1) \times K(S+1)}$ is a $K(S+1) \times K(S+1)$ matrix of zeros and $0_\epsilon = 0_{K((S+1)R-1)}$ is a $K((S+1)R-1) \times 1$ vector of zeros, and

$$\mathbf{\Phi_r} = \begin{pmatrix} \Phi_r & 0_{K\times K} & \dots & 0_{K\times K} \\ I_K & & 0_{K\times K} & 0_{K\times K} \\ & \ddots & & \vdots \\ 0_{K\times K} & & I_K & 0_{K\times K} \end{pmatrix} \quad \text{for } r \in \{1, \dots, R\}, \tag{5.25}$$

where $\Phi_r = 0_{K\times K}$ for all $r > P$. Equation (5.25) can then be replaced by

$$\mathbf{F_t} = \mathbf{H}(I_{K(S+1)R}L)\mathbf{F_t} + \epsilon_\mathbf{t} \quad \text{for } t \in \{1, \dots, T\}, \tag{5.26}$$

where L denotes the lag operator again. Rearranging the expression yields

$$\mathbf{F_t} = \left(I_{K(S+1)R} - \mathbf{H}(I_{K(S+1)R}L)\right)^{-1} \epsilon_\mathbf{t} \quad \text{for } t \in \{1, \dots, T\}, \tag{5.27}$$

which can be inserted into Equation (5.23), obtaining

$$y_t^* = \mathbf{\Lambda}\left(I_{K(S+1)R} - \mathbf{H}(I_{K(S+1)R}L)\right)^{-1} \epsilon_\mathbf{t} + \xi_t \quad \text{for } t \in \{1, \dots, T\}. \tag{5.28}$$

In the state-space form, the expression $\left(I_{K(S+1)R} - \mathbf{H}I_{K(S+1)R}L\right)^{-1}$ is the equivalent to $\Phi(L)^{-1}$ from Section 5.2.1 and must therefore be left coprime. This must also hold for all orthogonal transformations of the state-space system by a matrix $D \in O(K)$, which obtain as

$$y_t^* = \mathbf{\Lambda}(I_{(S+1)R} \otimes D)(I_{(S+1)R} \otimes D')\mathbf{F_t} + \xi_t \quad \text{for } t \in \{1, \dots, T\} \tag{5.29}$$

for the observation equation, and

$$(I_{(S+1)R} \otimes D')\mathbf{F_t} = (I_{(S+1)R} \otimes D')\mathbf{H}(I_{(S+1)R} \otimes D)(I_{(S+1)R} \otimes D')\mathbf{F_t}$$

$$+ (I_{(S+1)R} \otimes D')\epsilon_{\mathbf{t}} \quad \text{for } t \in \{1, \ldots, T\} \tag{5.30}$$

for the state equation. Thus Equations (5.29) and (5.30) represent the state-space system for arbitrary orthogonal transformations of the factors and loadings and hence the state-space system where the rotation problem remains unsolved.

5.2.3 The Ledermann Bound

Moreover, it must be ensured that the decomposition of the population covariance matrix Σ_y into a systematic and an idiosyncratic component is unique. This generally holds for $K \leq \varphi(N)$, where $\varphi(N)$ is the Ledermann bound, see e.g. Ten Berge and Sočan (2007). The Ledermann bound for the static factor model is

$$\varphi(N) = \frac{2N + 1 - \sqrt{8N + 1}}{2}, \tag{5.31}$$

see Ledermann (1937). This bound is derived from comparing the number of parameters in Σ_y, which is $\frac{N(N+1)}{2}$, or in the correlation matrix R_y, which is $\frac{N(N-1)}{2}$, with the number of parameters in the factor model. The number of parameters to be estimated is then $NK - \frac{K(K-1)}{2} + N$ or $NK - \frac{K(K-1)}{2}$, where Λ contains NK parameters. In the case of the correlation matrix, Λ determines the nonzero elements of Σ_y, so the σ_i^2 are no longer free parameters. The rotation problem introduces the matrix D, which reduces the number of uniquely determined parameters by $\frac{K(K-1)}{2}$.
The assumption that $N < T$ may not always hold, so taking the reduced rank of the covariance matrix Σ_y in such a case into account, there remain only $\frac{T(T+1)}{2}$ distinct elements, and the Ledermann bound accordingly changes to

$$\varphi(N, T) = \frac{2N + 1 - \sqrt{4(N + T)(N - T + 1) + 1}}{2} \quad \text{for} \quad T \leq N, \tag{5.32}$$

which nests the expression in Equation (5.31) for $T = N$.

Now consider the dynamic factor model with K factors, S lagged loadings matrices, P full persistence matrices in the factors, and Q diagonal persistence matrices describing the AR processes in the error terms. Thus there are $(S+1)NK+PK^2+(Q+1)N-\frac{K(K-1)}{2}$ parameters to be estimated. There is also much more information available from the data in terms of distinct parameters. For the autocorrelated data, the autocovariance matrices defined as

$$\Gamma_y(\tau) = \begin{pmatrix} \gamma_{1,1}(\tau) & \gamma_{1,2}(\tau) & \ldots & \gamma_{1,N}(\tau) \\ \gamma_{2,1}(\tau) & \gamma_{2,2}(\tau) & \ldots & \gamma_{2,N}(\tau) \\ \vdots & & \ddots & \vdots \\ \gamma_{N,1}(\tau) & \gamma_{N,2}(\tau) & \ldots & \gamma_{N,N}(\tau) \end{pmatrix} \quad \text{for } \tau \in \{0, T-1\}, \tag{5.33}$$

where $\gamma_{i,j}(\tau) = E[(y_{t,i} - E(y_{t,i}))(y_{t-\tau,j} - E(y_{t-\tau,j}))]$, can be determined, where $\Gamma_y(0) = \Sigma_y$. The block matrix of all autocovariance matrices

$$\mathbf{\Gamma} = \begin{pmatrix} \Gamma_y(0) & \Gamma_y(1) & \dots & \Gamma_y(T-1) \\ \Gamma_y(1)' & \Gamma_y(0) & \dots & \Gamma_y(T-2) \\ \vdots & \vdots & & \vdots \\ \Gamma_y(T-1)' & \Gamma_y(T-2)' & \dots & \Gamma_y(0) \end{pmatrix}, \tag{5.34}$$

which has dimension $NT \times NT$, therefore contains up to $\frac{TN(N+1)}{2}$ distinct parameters. The reduced rank case, however, must also be considered here. Hence, if $T \leq N$, the number of distinct parameters is $\frac{T^2(T+1)}{2}$. For the models considered in the empirical part of this Chapter the number of parameters is always far below this value, see also Kaufmann and Schumacher (2013).

5.3 Sampling Approach and Model Selection Criteria

To obtain estimates for the model parameters, I use a Gibbs sampling approach that leaves the rotation problem intentionally unsolved. The according unconstrained Gibbs sampler generally follows Carter and Kohn (1994), so the factors are obtained by forward-filtering backward-sampling, where the full conditional densities are obtained from the Kalman filter. For the high-dimensional models considered here with $N = 402$, however, the regular Kalman filter may incur numerical difficulties. One approach to deal with this is to use the sampler of Chan and Jeliazkov (2009), as in Kaufmann and Schumacher (2013), which allows to sample the factors in a single sweep without using a Kalman filter. An even higher numerical accuracy, however, can be achieved by simply replacing the regular Kalman filter by a square-root version described in Tippett et al. (2003), which is the approach taken in the following.

In the following, the analysis is limited to the case of $S = 0$, i.e. no loadings on the lagged factors are considered.[3] Hence Equation (5.5) is accordingly simplified, with $\Lambda = \Lambda_0$ and $F_t = f_t$.

As the Θ_q matrices are assumed to be diagonal, I define the following matrix containing the nonzero elements of all Θ_q for $q \in \{1, \dots, Q\}$:

$$\Theta = \begin{pmatrix} \theta_{1,1,1} & \dots & \theta_{Q,1,1} \\ \vdots & & \vdots \\ \theta_{1,N,N} & \dots & \theta_{Q,N,N} \end{pmatrix} \tag{5.35}$$

and denote the i^{th} row of this matrix as θ_i.

[3] Models with $S > 0$ are omitted from the analysis, as they generally turn out to be numerically difficult to handle. Other empirical analyses, e.g. Kaufmann and Schumacher (2012) and Kaufmann and Schumacher (2013) therefore also do not take these specifications into account.

Assuming a multivariate normal distribution for the innovations in the factors and idiosyncratic components in Equation (5.6), the likelihood of the data conditional on the parameters and the latent factors is

$$\mathcal{L}(\{y_t\}_{t=1}^T|\Lambda,\Sigma,\{\Phi_p\}_{p=1}^P,\{\Theta_q\}_{q=1}^Q,\{f_t\}_{t=1}^T) = (2\pi)^{-\frac{TN}{2}}\left|\Sigma'\right|^{-\frac{T}{2}} \times\prod_{t=1}^{T}\exp\left((\Theta(L)(y_t-\Lambda f_t))'\Sigma^{-1}(\Theta(L)(y_t-\Lambda f_t))\right), \tag{5.36}$$

where y_t denotes an $N\times 1$ vector of observations at time t and $\Theta(L)$ is the lag polynomial of the idiosyncratic error terms from Equation (5.8), see also Kaufmann and Schumacher (2013).[4]

5.3.1 The Unconstrained Gibbs Sampler for the Dynamic Factor Model

The Gibbs sampler for the dynamic factor model largely follows the setup of Otrok and Whiteman (1998) and Kose et al. (2003). The chosen prior distributions are independent, so their joint pdf is

$$\pi(\Lambda,\Sigma,\{\Phi_p\}_{p=1}^P,\{\Theta_q\}_{q=1}^Q) = \pi(\Lambda)\pi(\Sigma)\pi(\{\Phi_p\}_{p=1}^P)\pi(\Theta), \tag{5.37}$$

where the errors are uncorrelated across the variables, hence Σ is diagonal, so $\pi(\Sigma) = \prod_{i=1}^N \pi(\sigma_i^2)$ and Θ is also diagonal, so $\Theta = \prod_{i=1}^N \pi(\theta_i)$, with $\theta_i = (\theta_{i1},\ldots,\theta_{iQ})$.[5] The prior for $\Lambda = (\lambda_1' \quad \ldots \quad \lambda_N')'$ is assumed to factor as $\pi(\Lambda) = \prod_{i=1}^N \pi(\lambda_i)$, where the λ_i are assumed to follow normal distributions with mean zero and covariance $\underline{c}_i I_K$, with some constant $\underline{c}_i > 0$ for all $i \in \{1,\ldots,N\}$. The σ_i^2 are assumed to follow inverse Gamma distributions with hyperparameters $\underline{\alpha}_i$ and $\underline{\beta}_i$ for all $i \in \{1,\ldots,N\}$. As in Chapter 3, the prior distribution for $\{\Phi\}_{p=1}^P$ is chosen as uninformative, so $\pi(\{\Phi\}_{p=1}^P) \propto$ const. The prior distribution for the θ_i is chosen as a normal distribution with mean $\underline{\zeta}_i$ and covariance $\underline{\Psi}_i$ for all $i \in \{1,\ldots,N\}$. This choice of priors ensures invariance to orthogonal transformations. Throughout, the prior hyperparameters are chosen as $\underline{\alpha}_i = 1$, $\underline{\beta}_i = 1$, $\underline{c}_i = 1$, $\underline{\zeta}_i = 0$ and $\underline{\Psi}_i = I_Q$ for all $i \in \{1,\ldots,N\}$. By assumption, the innovations in the factors have mean zero. To fix the scaling of the factors and loadings, they are uncorrelated and have unit variance each, so

$$\pi\left(\{\epsilon_t\}_{t=1}^T\right) = \prod_{t=1}^{T}\pi(\epsilon_t) = \prod_{t=1}^{T}(2\pi)^{-\frac{K}{2}}\exp\left(-\frac{1}{2}\epsilon_t\epsilon_t'\right). \tag{5.38}$$

All model parameters and the factors are then iteratively sampled from their full conditional posterior distributions, for which a derivation is given in Appendix 5.A.

Convergence of the Gibbs sampler is monitored by Geweke's statistic, see Geweke (1991), for orthogonally invariant quantities, i.e. the 402 sums of squared loadings per variable, the 402 idiosyncratic error covariances, the P determinants of the persistence matrices in the factors

[4] Note that $\Theta(L)(y_t - \Lambda f_t) = \xi_t$ and $\xi_t \sim N(0,\Sigma)$.
[5] This is the Frisch case, discussed in Section 5.2.

Φ_p, and the $402Q$ persistence parameters in the idiosyncratic error terms. If in each group, convergence cannot be rejected for 90% of the monitored quantities, the sampler is assumed to have converged. After convergence, a sequence of 10,000 draws is used for inference.

In a subsequent step, the output of the unconstrained Gibbs sampler is postprocessed by the weighted orthogonal Procrustes (WOP) algorithm, as described in detail in Chapter 3. The mean of the postprocessed posterior sample then serves as the estimator. Accordingly, the estimates $\hat{\Lambda}$, $\hat{\Sigma}$, $\{\hat{\Phi}_p\}_{p=1}^P$ and $\hat{\Theta}$ are obtained.

5.3.2 Model Selection Criteria

In the following, I discuss two model selection criteria which are applicable in Bayesian factor analysis. Model selection for static factor models is discussed by Lopes and West (2004), who compare several criteria, to determine the number of factors, e.g. the Akaike information criterion (AIC), see e.g. Akaike (1974), two versions of the Bayesian information criterion (BIC), see e.g. Schwarz (1978), and the informational complexity (ICOMP) criterion by Bozdogan and Ramirez (1987), see also Bozdogan and Shigemasu (1998). These criteria are calculated based on maximum likelihood estimates of the model. As some of the model specifications considered in the application are rather complex, and the model is rich in parameters due to the dimension of the data set, where the number of cross-sections is about five times larger than the time dimension, finding the maximum likelihood estimates is rather cumbersome. I therefore only discuss two criteria that are particularly fit for a Bayesian analysis.

The first criterion is the deviance information criterion (DIC) by Spiegelhalter et al. (2002), which can be considered a generalization of the AIC, see e.g. Lee (2007) and requires the evaluation of the log likelihood for all draws from the Gibbs sampler, penalizing the term with the number of model parameters.[6] The criterion is hence

$$\text{DIC} = -\frac{2}{Z}\sum_{z=1}^{Z}\log\left(\mathcal{L}(Y|m,(\Lambda^{(z)},\Sigma^{(z)},\{\Phi_p^{(z)}\}_{p=1}^P,\{\Theta_q^{(z)}\}_{q=1}^Q),\{f_t^{(z)}\}_{t=1}^T)\right)+2d_m, \tag{5.39}$$

with $Y = (y_1, \ \ldots, \ y_T)'$, where z denotes the Gibbs sampler iteration, and Z denotes the length of the retained Gibbs sequence, and

$$d_m = NK + PK^2 + (Q+1)N - \frac{K(K-1)}{2} \tag{5.40}$$

denotes the number of model parameters, compare Section 5.2.2.

The second model selection criterion that suits the Bayesian estimation approach particularly well is the marginal likelihood. The marginal likelihood is a criterion used in Bayesian model

[6] The log likelihood value can be obtained from the Kalman filter, by evaluating the likelihood for the one-step ahead forecast error at every time point $t \in \{1,\ldots,T\}$.

selection based on the following argument, see e.g. Berger (1985), Chapter 3.5. Denoting a possible model choice by m, the probability to observe the data Y given model m is

$$p(Y|m) = \int p(Y|\vartheta, m)\pi(\vartheta|m)d\vartheta, \tag{5.41}$$

where ϑ is a vector containing all model parameters. Hence if two models m_0 and m_1 are compared, the marginal likelihood yields information about which model is more likely to have generated the observable data Y. From the marginal likelihoods, the Bayes factor can be calculated as

$$BF_{0,1} = \frac{p(Y|m_0)}{p(Y|m_1)}, \tag{5.42}$$

see e.g. Kass and Raftery (1995).

I evaluate the marginal likelihood by means of the Candidate's estimator, see e.g. Gilks et al. (1996), Chapter 10.4, which obtains the marginal likelihood from Chib's method, see Chib (1995), which in turn exploits the Candidate's formula from Besag (1989), which is simply a reformulation of Bayes' theorem and can be written for the dynamic factor model as

$$p(Y|m) = \frac{p(Y|m, (\Lambda, \Sigma, \{\Phi_p\}_{p=1}^P, \{\Theta_q\}_{q=1}^Q))\pi(\Lambda, \Sigma, \{\Phi_p\}_{p=1}^P, \{\Theta_q\}_{q=1}^Q|m)}{p(\Lambda, \Sigma, \{\Phi_p\}_{p=1}^P, \{\Theta_q\}_{q=1}^Q|m, Y)}, \tag{5.43}$$

where m denotes the chosen model, the first expression in the numerator is the likelihood of the data under model m, the second expression is the prior under model m. The expression in the denominator, which is the joint posterior density of all model parameters, is then obtained applying Chib's method to the dynamic factor model. Noting that this expression is the prediction density from Besag (1989), the marginal likelihood expressed in Equation (5.43) measures how well the data is predicted by the choice of model m.

As Equation (5.43) holds for any $(\Lambda, \Sigma, \{\Phi_p\}_{p=1}^P, \{\Theta_q\}_{q=1}^Q)$, insert the Bayesian parameter estimates, i.e. the posterior means, $(\Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, \{\Theta_q^*\}_{q=1}^Q)$. Next, take the log of the resulting expression and obtain the estimated marginal density

$$\begin{aligned}\log(\hat{p}(Y|m)) = &\log(p(Y|m, (\Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, \{\Theta_q^*\}_{q=1}^Q))) \\ &+ \log(\pi(\Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, \{\Theta_q^*\}_{q=1}^Q|m)) \\ &- \log(\hat{p}(\Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, \{\Theta_q^*\}_{q=1}^Q|m, Y)),\end{aligned} \tag{5.44}$$

where the last expression is the log of a Monte Carlo estimate of the representation as conditional densities

$$\begin{aligned}p(\Lambda, \Sigma, \{\Phi_p\}_{p=1}^P, \{\Theta_q\}_{q=1}^Q|m, Y) = &\, p(\Lambda|m, Y)p(\Sigma|\Lambda, m, Y)p(\{\Phi_p\}_{p=1}^P|\Lambda, \Sigma, m, Y) \\ &\times p(\{\Theta_q\}_{q=1}^Q|\Lambda, \Sigma, \{\Phi_p\}_{p=1}^P, m, Y),\end{aligned} \tag{5.45}$$

at $(\Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, \{\Theta_q^*\}_{q=1}^Q)$, which is

$$\hat{p}(\Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, \{\Theta_q^*\}_{q=1}^Q | m, Y) = \hat{p}(\Lambda^* | m, Y)\hat{p}(\Sigma^* | \Lambda^*, m, Y)\hat{p}(\{\Phi_p^*\}_{p=1}^P | \Sigma^*, \Lambda^*, m, Y) \times \hat{p}(\{\Theta_q^*\}_{q=1}^Q | \{\Phi_p^*\}_{p=1}^P, \Sigma^*, \Lambda^*, m, Y), \tag{5.46}$$

where each of the four expressions on the right hand side is separately evaluated, where the factors are sampled as augmented parameters in each evaluation step.

An approximation of the first expression on the right hand side of Equation (5.46) obtains directly from the Gibbs output as

$$\hat{p}(\Lambda^* | m, Y) = \frac{1}{Z} \sum_{z=1}^{Z} p(\Lambda^* | \Sigma^{(z)}, \{\Phi_p^{(z)}\}_{p=1}^P, \{\Theta_q^{(z)}\}_{q=1}^Q; \{f_t^{(z)}\}_{t=1}^T, m, Y), \tag{5.47}$$

where Z denotes the length of the retained sequence from the Gibbs sampler and it is simply required to evaluate the conditional probability of Λ^* for the draws of Σ, $\{\Phi_p\}_{p=1}^P$ and $\{\Theta_q\}_{q=1}^Q$.

The second expression conditions on Λ^*, so it cannot use the Gibbs output, but requires running the Gibbs sampler anew, with Λ fixed at Λ^*, hence obtaining draws of the augmented factors from the forward-filtering backward-sampling approach

$$\{f_t^{(z)} | \Lambda^*, \Sigma^{(z-1)}, \{\Phi_p^{(z-1)}\}_{p=1}^P, \{\Theta_q^{(z-1)}\}_{q=1}^Q\}_{t=1}^T \tag{5.48}$$

and sampling from the full conditional densities

$$f(\Sigma^{(z)} | \Lambda^*, \{\Phi_p^{(z-1)}\}_{p=1}^P, \{\Theta_q^{(z-1)}\}_{q=1}^Q; \{f_t^{(z)}\}_{t=1}^T, m, Y), \tag{5.49}$$

$$f(\{\Phi_p^{(z)}\}_{p=1}^P | \Lambda^*, \Sigma^{(z)}, \{\Theta_q^{(z-1)}\}_{q=1}^Q; \{f_t^{(z)}\}_{t=1}^T, m, Y), \tag{5.50}$$

$$\text{and} \quad f(\{\Theta_q^{(z)}\}_{q=1}^Q | \Lambda^*, \Sigma^{(z)}, \{\Phi_p^{(z)}\}_{p=1}^P; \{f_t^{(z)}\}_{t=1}^T, m, Y) \tag{5.51}$$

for $z \in \{1, \ldots, Z\}$, which yields the estimate

$$\hat{p}(\Sigma^* | \Lambda^*, m, Y) = \frac{1}{Z} \sum_{z=1}^{Z} p(\Sigma^* | (\{\Phi_p^{(z)}\}_{p=1}^P | \Lambda^*), (\{\Theta_q^{(z)}\}_{q=1}^Q | \Lambda^*); (\{f_t^{(z)}\}_{t=1}^T | \Lambda^*), m, Y). \tag{5.52}$$

The third expression conditions on Λ^* and Σ^* and requires an additional run of the Gibbs sampler with Λ fixed at Λ^* and Σ fixed at Σ^*. Again, the draws of the factors are obtained from the forward-filtering backward-sampling approach as

$$\{f_t^{(z)} | \Lambda^*, \Sigma^*, \{\Phi_p^{(z-1)}\}_{p=1}^P, \{\Theta_q^{(z-1)}\}_{q=1}^Q\}_{t=1}^T, \tag{5.53}$$

and the draws for the remaining parameters are obtained from the full conditional densities

$$f(\{\Phi_p^{(z)}\}_{p=1}^P | \Lambda^*, \Sigma^*, \{\Theta_q^{(z-1)}\}_{q=1}^Q; \{f_t^{(z)}\}_{t=1}^T, m, Y), \tag{5.54}$$

$$\text{and} \quad f(\{\Theta_q^{(z)}\}_{q=1}^Q | \Lambda^*, \Sigma^*, \{\Phi_p^{(z)}\}_{p=1}^P; \{f_t^{(z)}\}_{t=1}^T, m, Y) \tag{5.55}$$

for $z \in \{1, \dots, Z\}$, which yields the estimate

$$\hat{p}(\{\Phi_p^*\}_{p=1}^P | \Lambda^*, \Sigma^*, m, Y) = \frac{1}{Z} \sum_{z=1}^{Z} p(\{\Phi_p^*\}_{p=1}^P | (\{\Theta_q^{(z)}\}_{q=1}^Q | \Lambda^*, \Sigma^*); (\{f_t^{(z)}\}_{t=1}^T | \Lambda^*, \Sigma^*), m, Y). \tag{5.56}$$

Eventually, the fourth expression conditions on Λ^*, Σ^* and $\{\Phi_p^*\}_{p=1}^P$ and requires one final run of the Gibbs sampler with Λ fixed at Λ^*, Σ fixed at Σ^* and $\{\Phi_p\}_{p=1}^P$ fixed at $\{\Phi_p^*\}_{p=1}^P$. The draws of the factors are again obtained from forward-filtering backward-sampling as

$$\{f_t^{(z)} | \Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, \{\Theta_q^{(z-1)}\}_{q=1}^Q\}_{t=1}^T, \tag{5.57}$$

and the draws for $\{\Theta_q\}_{q=1}^Q$ are obtained from the full conditional density

$$f(\{\Theta_q^{(z)}\}_{q=1}^Q | \Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P; \{f_t^{(z)}\}_{t=1}^T, m, Y) \tag{5.58}$$

for $z \in \{1, \dots, Z\}$, which yields the estimate

$$\hat{p}(\{\Theta_q^*\}_{q=1}^Q | \Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P, m, Y) = \frac{1}{Z} \sum_{z=1}^{Z} p(\{\Theta_q^*\}_{q=1}^Q | (\{f_t^{(z)}\}_{t=1}^T | \Lambda^*, \Sigma^*, \{\Phi_p^*\}_{p=1}^P), m, Y). \tag{5.59}$$

Note that in the three conditional Gibbs sampler runs, Λ is set to Λ^*, so there is no rotation problem in the output of the sampler, and hence, no postprocessing is necessary for the additionally sampled three sequences.

5.4 Data Description and Model Selection

The analyzed data set consists of $N = 402$ time series at monthly frequency, one for each county, or NUTS-3 region, in Germany. The counties are listed in Table 5.1. The data covers $T = 82$ months from January 2007 until October 2013. Thus the time series are rather short, covering roughly one business cycle. This is due to the labor market reforms of 2005, which brought substantial legislative changes, which result in figures before and after this time point not being comparable. Moreover, figures from 2005 and from parts of 2006 may be unreliable in some cases due to measurement issues that occurred in the initial phase of the newly established administrative structure. The small T makes it difficult to investigate on convergence or decoupling, however, it suffices to identify common patterns. The question of convergence or decoupling may be partially answered by comparing the findings of this chapter to those of Boysen-Hogrefe and Pape (2011), who find an initial dichotomy between former East and West Germany in the data up until 2004, which is greatly reduced until 2010.

The original series of absolute unemployment figures are denoted $x_{i,t}$ with $i \in \{1, \dots, 402\}$ denoting the county and $t \in \{1, \dots, 82\}$ denoting the month in the sample. I refrain from the use of unemployment rates, which may partially conceal the labor market dynamics, depending both on the number of employed and unemployed persons, so if these two figures evolve similarly, the unemployment may be unchanged, despite substantial changes in both the number of employed and unemployed persons. Moreover, employed persons are counted with respect to the place of work, whereas unemployed persons are counted with respect to the place of residence, so unemployment rates automatically bring in the question of labor mobility. Eventually, the number of employed persons is generally only available at quarterly intervals on county level. To circumvent these questions, which are beyond the scope of the chapter, I use absolute unemployment figures, which are available at monthly frequency.

Regarding the dimension of the factor model, it is important to note that the time dimension T is very small compared to the number of time series N. This is not uncommon in macroeconomic factor models, however, as the data set analyzed in Kose et al. (2003) contains 3 series for 60 countries, hence $N = 180$, while the time dimension is only $T = 30$. For cross-sectional factor analysis, this has been called the "small N large p" case by West (2003), who proposes the use of sparse factor models in this case. Sparse factor models have successfully been applied to large macroeconomic data sets by Kaufmann and Schumacher (2012) and Kaufmann and Schumacher (2013). Nonetheless, my analysis starts with a model having a full loadings matrix, while in a subsequent step, the two-step approach for sparse factor analysis proposed in Chapter 4 is applied.

5.4.1 Preprocessing the Data for the Analysis

The absolute unemployment figures depend strongly on the size of the considered county, and moreover, the series may not be stationary. The left panel of Figure 5.1 shows the distribution of the first-order autocorrelations of the 402 series. The values are generally close to 1, so most series seem in fact non-stationary. Therefore, following Boysen-Hogrefe and Pape (2011), I transform all series into growth rates, as

$$y_{i,t} = \frac{x_{i,t} - x_{i,t-1}}{x_{i,t-1}} \quad \text{for} \quad t \in \{2, \dots, 82\}, \tag{5.60}$$

for the series y_i, which each have length $T = 81$ now. The first-order autocorrelations for the growth rates, shown in the right panel of Figure 5.1, are generally below 0.5 in magnitude, so stationarity can be assumed to hold for the growth rates.

Moreover, unemployment rates generally contain a seasonal component, which should be removed before proceeding. To do so, approaches like X-11 ARIMA or X-12 ARIMA are commonplace in applications nowadays, see e.g. Ladiray and Quenneville (2001) or Findley et al. (1998). The series y_i for $i \in \{1, \dots, 402\}$ are all seasonally adjusted by X-12 ARIMA.[7]

[7] The seasonal adjustment may also be implemented before transforming the series into growth rates, but the results do not change substantially.

The left panel of Figure 5.2 shows the average monthly growth of unemployment over the entire sample period, calculated as the geometric mean. In most regions, unemployment is decreasing over the sample period on average, with the exception of Nordfriesland, Wittmund, the Mosel region, and several counties in Bavaria, most notably in the eastern part. The exceptions coincide with the regions showing the most pronounced seasonal pattern. The right panel of Figure 5.2 shows the amplitude of the seasonal components for all counties, which has a correlation of 0.66 with the average growth rate in the left panel. Investigating aspects of seasonal unemployment, Karr (1983) finds that out of the group of people that are regularly seasonally unemployed over a period of at least five years, 65% live in Bavaria. These results are 30 years old and were obtained using data for West Germany only, but the pattern persists in more current data, ranging from 1998 to 2007, which is used in the study by Schanne et al. (2010), and is obviously still present here.

5.4.2 Model Selection

To find an appropriate model for the data, 105 specifications are compared. The number of factors varies between 1 and 7, paired with 0 to 4 lags in the factor structure, and with 0 to 2 lags in the idiosyncratic component of each data series.

Table 5.2 reports the results of the DIC for the 105 considered model specifications. The minimum is reached for a model with $K = 1$ factor, $P = 1$ lag in the factor, and no dynamic structure in the idiosyncratic components, i.e. $Q = 0$. Table 5.3 reports the log marginal likelihood, where the maximum is reached for a model with $K = 7$ factors, $P = 0$ lags in the factors and no dynamic structure in the idiosyncratic components. These results are quite different from each other, so they are separately dealt with in the following.

Using the log marginal likelihood values, the Bayes factor from Equation (5.42) for the comparison of two models m_0 and m_1 can be obtained as

$$BF_{0,1} = \exp(\log(p(Y|m_0)) - \log(p(Y|m_1))), \tag{5.61}$$

so a comparison of the model with $K = 7$, $P = 2$ and $Q = 0$ to the one with $K = 7$, $P = 1$ and $Q = 0$ yields a log Bayes factor of 32.49. Even though this says that the first model is more than thirty times more likely to be the true one than the second, numerical uncertainty may imply that they are barely distinguishable from each other. The small T is also an issue here, as e.g. Robert (2014) notes in a different context that the distribution of the Bayes factors for rather small sample sizes often does not allow to discriminate well between models. Thus even though the log Bayes factor is 32.35 in favor of the model with $K = 7$ non-autocorrelated factors when comparing it to the model with $P = 1$, and 64.84 when comparing it to the model with $P = 1$, the results should be treated with caution. Due to the MCMC approach, numerical variation must be accounted for when comparing the criteria. The extent of the numerical variation from the Gibbs sampling approach is unknown, but can be analyzed by repeating the evaluation of the log marginal likelihood. To find out whether the realizations of

the log marginal likelihoods are merely coincidental, the log Bayes factors for the comparison between the three models with the highest log marginal likelihood values are each re-estimated 20 times. To find out about the numerical variation in the log Bayes factors, I create a sample of size 1,000 for each log Bayes factor by bootstrapping from the 20 estimates for the log marginal likelihood for each model. A histogram of the distributions of the log Bayes factors is shown in Figure 5.3. The histograms appear to be in favor of the model with $K = 7$ factors and $P = 0$ lags in the factor structure compared to the models with dynamic factors, whereas the distinction between the model with $K = 7$ factors and $P = 1$ or $P = 2$ lags in the factors, respectively, is not so clear.

Instead of choosing the model with $P = 0$ straightaway, I apply a second diagnostic tool that can be used to determine the lag order in the estimated factors. The factors are estimated by the posterior mean of the sampled augmented factors from the unconstrained sampler, which are transformed by the same sequence of orthogonal matrices that is used to transform the sequence of Λ and $\{\Phi_p\}_{p=1}^P$ in the WOP approach, and are denoted as $\{\hat{f}_t\}_{t=1}^T$, or, in matrix form, as $\hat{F}$. The autocorrelation functions (ACFs) of the factor estimates may reveal whether the lag order has been properly chosen by looking at the Bayes factor. I start considering the model with $K = 7$ static factors and perform a rotation of the estimated factors and the corresponding loadings, to mimic the outcome of a factor model based on Principal Components Analysis (PCA), i.e. the first factor explains most of the variation, the second factor explains most of the remaining variation, and so forth. This approach roughly follows Kose et al. (2003), who perform an orthogonalization of the sampled factors. In addition, the factors are also scaled here, using a transformation applied in the sampling approach of Kaufmann and Schumacher (2012).

Consider the estimates obtained for the latent factors, and calculate their empirical covariance matrix $\hat{\Sigma}_f$. Then perform a spectral decomposition of this empirical covariance[8] as

$$\hat{\Sigma}_f = VDV', \tag{5.62}$$

where V contains the eigenvectors of $\hat{\Sigma}_f$ as row vectors, and D is a diagonal matrix of eigenvalues. Let

$$H = VD^{\frac{1}{2}}, \tag{5.63}$$

and use this expression to transform

$$\{\hat{f}_t^*\}_{t=1}^T = \{H\hat{f}\}_{t=1}^T, \tag{5.64}$$

$$\hat{\Lambda}^* = \hat{\Lambda}H', \quad \text{and} \tag{5.65}$$

$$\{\Phi_p^*\}_{p=1}^P = \{H'\Phi_p H\}_{p=1}^P. \tag{5.66}$$

[8]Note that orthogonal unit scale factors could also be obtained by using a Cholesky decomposition, but the spectral decomposition has the advantage of automatically ordering the factors by their contribution to the total explained variation, as in PC factor analysis.

The transformation matrix H can likewise be used to transform all elements of the sampled Gibbs sequences, e.g. to obtain highest posterior density intervals (HPDIs) for the accordingly transformed parameters.

Figure 5.4 shows the ACFs for the resulting factors. The first factor is clearly autocorrelated, and for the second factor, there is at least some evidence of autocorrelation. The ACFs of the remaining factors do not indicate any autocorrelation. Thus, I consider next the model with $K = 7$ dynamic factors, where $P = 1$. Figure 5.5 shows the ACFs of the residuals from filtering the factors with the estimated Φ, i.e. $\hat{f}_t - \hat{\Phi}\hat{f}_{t-1}$ for $t \in \{2, \ldots, T\}$. The obvious autocorrelation is gone there, so the model with $K = 7$ and $P = 1$ seems appropriate.

Furthermore, I double-check whether the model selection by the estimated log marginal likelihood was right with its preference for $Q = 0$ by looking at the ACFs of the residuals. Almost all of them show no suspicious pattern indicating autocorrelation, except for the 16 series shown in Figure 5.6, where more than three out of the 50 autocorrelation coefficients exceed the approximate 95% significance bounds. Except for very few series, which amount to less than 4% of all series, the idiosyncratic terms do not appear autocorrelated, so I use the model with $Q = 0$ in the following.

As mentioned above, the DIC favors a model with only $K = 1$ factor. Since the number of parameters d_m grows quickly in K due to the large number of series in the data set, additional factors are strongly penalized by the criterion, which may explain the choice of $K = 1$.[9] It can be argued, however, that the model with $K = 7$ factors should be chosen. Apart from possibly using up too many degrees of freedom and diminishing forecasting performance, overestimating the number of factors still yields consistent estimates, while underestimating the number of factors does not. This has been shown for Principal Components (PC) factor analysis e.g. by Stock and Watson (2002a). For the Bayesian approach, the same argument holds by the fact that by erroneously specifying K too large, introducing K_{sp} spurious factors, the posterior estimate of Λ obtained from the WOP approach is an orthogonal transformation of a matrix where the loadings on the spurious factors on all variables are close to zero. As the estimate of Λ from the WOP approach can be orthogonally transformed to facilitate interpretation, see Chapter 3, these K_{sp} columns and the according spurious factors can be restored by an appropriate orthogonal transformation, and accordingly, the K non-spurious factors are restored. On the same grounds, the approach in Chapter 4 can be used to identify spurious factors, which is done in Section 5.6.

5.5 WOP Estimates and the Relation to PC Factor Analysis

In the following, I describe the results of the selected model with $K = 7$, $P = 1$ and $Q = 0$ estimated by means of the unconstrained sampler and postprocessed by the WOP procedure.

[9] Five replications of the estimation for the three models with the smallest values for the DIC yield the same ordering of the three models every time, i.e. the smallest DIC values are reached for the model with $K = 1$, $P = 1$ and $Q = 0$.

Figure 5.7 shows the median, 68%, 90% and 95% highest posterior density intervals (HPDIs) for the factor estimates transformed as in Equation (5.64). For all three intervals, both the average width and the standard deviation is strictly monotonically increasing. The HPDIs of the first factor are by far the most narrow ones, indicating that it is the factor for which the most precise estimates can be obtained.

5.5.1 Numerical Properties and Comparison to PC Factor Analysis

To evaluate the numerical variation in the estimates, I perform 20 repeated estimations of the selected model, where I use different seeds for the random number generator. Figure 5.8 shows the results plotted on top of each other. Just as in the dynamic application in Chapter 3, where the estimates of the factors and parameters are transformed in a different fashion, the Monte Carlo error is negligible.[10] Moreover, the corresponding factors from PC factor analysis are shown. The correlation between the first scaled factor from the Bayesian estimation approach and that of the first principal component is 0.9997, that of the second and third with their PC counterparts is 0.9978 and 0.9951, respectively.[11] This is an interesting result, since PC factor analysis is not only the most simple method to extract orthogonal latent factors from a data set, but also the one that ensures that the factors are chosen in such a way that they explain as much of the variation in the data as possible. PC factor analysis is shown to be an appropriate tool if an approximate factor structure is present in the data by Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986, 1988, 1993). Based on the PC approach for approximate static factor models, dynamic approaches are developed e.g. by Stock and Watson (2002a,b) and Doz et al. (2011), while other model extensions working iteratively include e.g. the factor-augmented vector-autoregressive (FAVAR) model by Bernanke et al. (2005). All these methods, however, rely on the initial PCA results obtained under the assumption of an approximate factor structure, which makes them fundamentally different from the Bayesian approaches e.g. by Kose et al. (2003) or Kose et al. (2012, 2008). Despite this difference, initial estimates from PC factor analysis are sometimes used as a starting point for Bayesian factor analysis, as in the hierarchical dynamic factor model approach proposed by Moench et al. (2009). The similarity between the transformed results from the Bayesian approach with WOP and a simple PC factor analysis observed here, however, indicates that a conformable model setup in terms of the number of factors in Bayesian factor analysis with WOP can get very close to the results of PC factor analysis.[12] In particular, the large correlation of the transformed WOP estimates with the PC factor estimates indicates that the

[10] The average MC error over all factors at all observed time points is 0.0076 and never exceeds 0.05, with an average of the first factor of only 0.0007 and of the seventh factor of 0.0146.

[11] For the following four factors, the correlation still exceeds 0.96. The correlation figures are in the same range for the estimated factor loadings.

[12] And vice versa, which may raise the question whether PCA based estimation approaches provide sufficiently good estimates to render Bayesian factor analysis unnecessary. It must be noted, however, that the model setup here is a very simple one, with $P = 1$ and $Q = 0$. Bayesian factor analysis allows e.g. to incorporate cross-correlation in the errors, to incorporate prior information and to estimate sparse loadings matrices. Moreover, in this chapter, the posterior densities obtained from the WOP approach are important, as they serve to estimate a sparse loadings structure for the model, see Section 5.6.

estimates obtained from the WOP approach fare well with respect to the share of explained variation.

5.5.2 Economic Interpretation of the Results

Next, I investigate the economic meaning of the factor estimates. Note first that the seven factors discussed so far have been constructed on the basis of their statistical properties, i.e. they were rotated in order to be perfectly uncorrelated, which produced factors closely resembling the first seven principal components. After scaling each factor to unit variance, the factors share the principal components' property of being ordered by explanatory power, starting with the factor explaining the largest share of the variation in the data. These factors are most likely not economically meaningful: If there are indeed seven shocks or cycles driving the observed changes in unemployment, they are most likely not identical to the principal components. Therefore I perform an orthogonal transformation of the factors such that at least some of them become interpretable.

As discussed earlier, the unemployment figures should reflect the business cycle, or, more generally, economic conditions. As GDP data is only available at quarterly frequency, I impute the missing data by interpolation and additionally use the ifo business climate indicator and business expectations indicator instead. Whereas the absolute correlation with the GDP does not exceed 0.8 for any orthogonal transformation, for the business climate, the highest absolute value of the contemporaneous correlation is 0.8503, whereas for the business expectations, the highest absolute correlation is obtained for a two months lag, reaching 0.8824.[13] Assuming that the business climate indicator is a better proxy for the business cycle than the lagged business expectations, I choose the former measure and perform a rotation of the factors, loadings and persistence parameters that yields a contemporaneous correlation of -0.8503 with the business climate indicator. The first factor in the model is therefore denoted as the *business climate factor* in the following. The average share of variation that is explained by all factors combined is 40.87%, and the business climate factor alone is able to explain 18.23% of the total variation on average.

The remaining six factors, which explain on average another 22.64% of the variation per series, are sequentially rotated to maximize their kurtosis. Following the concept of independent component analysis (ICA), see e.g. Hyvärinen et al. (2001), this ensures that the factors are dominated by few shocks, which may be traced back to particular signals, or particular events. As the factors have mean zero and unit variance, the kurtosis for the k^{th} factor obtains as

$$\kappa_k = \frac{1}{T-2} \sum_{t=1}^{T} f_{kt}^4. \tag{5.67}$$

While the kurtosis for the business cycle factor is 3.6975, which resembles the value of 3 that holds for normally distributed data, the remaining rotated factors reach substantially higher

[13] All correlations have a negative sign, which is economically plausible for a comparison of a measure of the business climate or business expectations with a measure of unemployment.

values of up to 28.7627, shown in Table 5.4, which also reports the average explanatory power of each of the factors. Moreover, it can be seen that the business climate factor does not explain the largest part of the variance for all counties, but only for 221 of them, whereas the remaining counties are dominated by other factors, particularly the fourth factor. Figure 5.9 shows the HPDIs for the according factors, and Figure 5.10 shows the corresponding loadings. The signs of the factors and loadings have been adjusted such that the average loading on each factor is positive. Table 5.5 reports the number of positive loadings per factor, which is usually rather high. If a factor has exclusively, or almost exclusively, positive loadings, this indicates that it represents some general cycle and is not merely coincidental. For the business climate factor, as well as for the fourth, fifth and seventh factor, more than 90% of the loadings have a positive sign.

The business climate factor's strong negative correlation with the ifo business climate indicator can be seen in the first panel of Figure 5.9. The first panel of Figure 5.10 shows that high loadings on this factor can be found in in the South and the West, where Märkischer Kreis, Esslingen and Heilbronn reach the highest values, whereas small and negative loadings occur mostly in the North and the East. Table 5.6 shows the average loadings per country (Bundesland). Bavaria and Baden-Württemberg, which are generally considered the economically most successful countries (Bundesländer), reach the highest average loadings, followed by North Rhine-Westphalia, Hamburg and Saarland. With regard to the business cycle exposure, there is thus a dichotomy, however not as pronounced as found by Boysen-Hogrefe and Pape (2011). Instead of a division between West and East Germany, there is rather a division between the North-East and the South-West.

As shown in the second panel of Figure 5.9, the second factor is dominated by a single positive shock in January 2012. The factor explains only 1.45% of the total variation per series on average and reaches the largest values in some regions in the eastern part of Bavaria, which are also most strongly exposed to seasonal unemployment, as discussed in Section 5.4.[14] The third factor shown in the third panel of Figure 5.9 features two positive shocks, in August 2007 and August 2012, and two negative ones, in August 2009 and August 2010. A look at the third panel in Figure 5.10 reveals that its predominantly positive loadings are concentrated on four regions, namely Lower Saxony, Saxony, Saxony-Anhalt and Thuringia. This factor only explains 2.58% of the variation on average, but almost 11% on average in Lower Saxony and between 6 and 7% in Saxony, Saxony-Anhalt and Thuringia. Even though both the time pattern in the third panel of Figure 5.9 and the regional pattern in the third panel of Figure 5.10 looks very systematic, a clear-cut explanation is not easily found.

The fourth factor, displayed in the fourth panel of Figure 5.9, shows overall more variation than the previous two and is dominated by a large negative shock in January 2008. In fact, there was a sharp decrease in unemployment in that month, preceding the financial crisis. The factor does not show substantial correlation with any of the ifo indicators or the GDP growth at any lag. The loadings on this factor are positive for all counties, and the factor explains

[14] This factor is not easy to interpret and might even represent some artifacts of the deseasonalization.

on average 11.56% of the variation per series. The average loadings shown in Table 5.6 are very similar across the countries (Bundesländer), so it appears to capture some general labor market dynamics that are not directly related to the business cycle. The fifth factor, shown in the fifth panel of Figure 5.9 and explaining 2.89% of the variation on average, is dominated by two negative shocks in January 2009 and April 2010, and a positive one in January 2011. It is most prominent in Hamburg, Schleswig-Holstein, Mecklenburg-Vorpommern and Brandenburg. The sixth factor features some fluctuations during the financial crisis and has substantial negative loadings only in Hesse and Rhineland-Palatinate, where it explains on average 4% of the variation, whereas the average explained variation for all of Germany is only 1.56%.[15] The seventh factor has an average explanatory power of 2.61% and does not contain any remarkable shocks. Altogether, maximizing the kurtosis in fact yields factors that are dominated by few shocks and that are to some extent interpretable. On the other hand, some of the spikes in the factors may be due to errors in the deseasonalization. Moreover, the concept of kurtosis maximization is quite arbitrary and can easily be replaced by a rotation of the solution subject to other criteria, such as Varimax or Quartimax.

5.6 Estimating a Sparse Model

As the Gibbs sampler output postprocessed by WOP allows to construct HPDIs for the loadings parameters, such intervals can be used to decide whether a loadings parameter is different from zero or not, depending on whether zero is included in the corresponding $1 - \alpha$ HPDI. If zero is included in the interval, the parameter can be considered "insignificant" in a Bayesian sense, see e.g. Hoff (2009). When applied to the loadings matrix in a factor model, this implies that some of the entries are zero, and the matrix has thus a sparse structure. In this section, I use the result for the full model to construct univariate HPDIs.

Moreover, the results obtained by means of the WOP approach reported in Section 5.5 can be arbitrarily orthogonally transformed. In the previous section I therefore chose a representation that maximizes the absolute correlation of the first factor with the ifo business climate indicator and otherwise maximizes the kurtosis of the respective factors. In this section, I therefore also apply the model-based approach from Chapter 4 that provides a sparse model representation. The approach is based on multivariate HPDIs for the loadings parameters, which are orthogonally transformed to find a parsimonious loadings structure.

5.6.1 A Sparse Model with Univariate HPDIs

The findings based on univariate HPDIs for the maximum kurtosis solution from Section 5.5 are shown in Table 5.7 for different lengths of the intervals in terms of α. Using a shorter interval will result in the zero less likely to be included. Based on a 68% HPDI, the zero is included in the HPDIs for 35 loadings on the first factor, based on the 90% HPDI, 56 intervals

[15] In Hesse, the financial sector plays an important role, so this finding may indicate its relation to the sixth factor.

contain the zero, and based on the 95% HPDI, 71 intervals contain the zero. Except for the business cycle factor and the fourth factor, the 95% HPDIs contain the zero for more than half of the loadings parameters, indicating that the corresponding factors are in fact restricted to specific regions.

Next, I estimate a factor model which imposes the sparse loadings structure identified using the 95% HPDI, i.e. I perform a Bayesian confirmatory factor analysis. Figure 5.11 shows the HPDIs for these factors. The factors resemble those for the full model shown in Figure 5.9 - the correlations with their counterparts from the full model amount to 0.9379, 0.9569, 0.8819, 0.9664, 0.8350, 0.9585 and 0.7465, respectively. The HPDIs of the factors are narrower than previously, as shown in Table 5.8. Regarding the loadings, the results are generally similar to those from the full model, as can be seen by comparing Figure 5.12 to Figure 5.10. Some entries of the loadings matrix that were previously close to zero are now set to zero, whereas others are larger in absolute terms. The fourth factor stands out, however: Instead of positive loadings throughout, it now has some positive and some negative loadings, which are generally not too far from zero. Table 5.7 shows that for the first, fifth and seventh factor, almost all nonzero loadings have a positive sign now. The fourth factor, which had almost exclusively loadings with positive signs in the full model, has now 99 loadings with negative signs and no longer appears to be a general factor.

As can be seen from Figure 5.13, the difference in total explanatory power of the model is overall small.[16] A look at Table 5.9, however, shows that the regional patterns are more pronounced in the sparse model than in the full model. The first factor has large positive loadings in North Rhine-Westphalia, Hamburg, Baden-Württemberg, Bavaria and Saarland, and to a smaller extent in Rhineland-Palatinate. The third factor has large positive loadings in Lower Saxony, Saxony, Saxony-Anhalt and Thuringia. The fourth factor has the largest positive loadings in and around Berlin. The fifth factor is pronounced in Schleswig-Holstein, Hamburg and Mecklenburg-Vorpommern, and to a smaller extent for Brandenburg. If regional clustering beyond country (Bundesland) borders is considered, the sixth and seventh factor likewise show some systematic behavior: The former has again substantial negative loadings in parts of Hesse and Rhineland-Palatinate, and the latter has its positive loadings in regions most affected by seasonal unemployment.

5.6.2 A Sparse Model with Multivariate HPDIs

Having estimated a sparse model in a rather arbitrary fashion based on the univariate HPDIs of the factor loadings, which helped to identify some regional patterns, I next apply the approach from Chapter 4. It uses the fact that the K-variate posterior distributions for each row of the loadings matrix λ_i for $i \in \{1, \dots, N\}$ are elliptical. Thus instead of univariate HPDIs, multivariate HPDIs, or highest posterior density ellipsoids (HPDEs) can be constructed. Due to the orthogonal invariance of the WOP output, the obtained sample from the posterior density of the parameters is just one of infinitely many possible orthogonal transformations.

[16] Only Offenbach and Burgenlandkreis do not load on any of the factors in the sparse model.

The aim of the approach from Chapter 4, conversely, is to find an orthogonal transformation of the HPDEs that allows for a parsimonious loadings structure with respect to a predefined criterion. For instance, the most parsimonious loadings structure, i.e. the one with the largest number of zero entries, may be of interest.

The HPDE for each λ_i can be constructed from the obtained sample by first calculating the Mahalanobis distance of each sampled $\lambda_i^{(z)}$ from the sample mean $\frac{1}{Z}\sum_{z=1}^{Z}\lambda_i^{(z)} = \widehat{E(\lambda_i)}$ by

$$d_i^{(z)} = (\lambda_i^{(z)} - \widehat{E(\lambda_i)})'\hat{\Sigma}_i^{-1}(\lambda_i^{(z)} - \widehat{E(\lambda_i)}), \tag{5.68}$$

where $\hat{\Sigma}_i = \widehat{\mathrm{Cov}(\lambda_i)}$. This calculation is performed for every $i \in \{1, \ldots, N\}$.

Next, depending on the desired width of the HPDEs, which is $1-\alpha$, the outermost $\lfloor \alpha Z \rfloor$ points for every λ_i, i.e. the points located at the largest Mahalanobis distance from the mean, are discarded. Note that an orthogonal transformation of the entire sample leaves the Mahalanobis distance of each point unchanged. In order to find a parsimonious representation, the HPDEs are then jointly orthogonally transformed to maximize a criterion that is chosen in accordance with the desired structure in Λ. The criterion may simply count the number of elements of Λ for which the zero is simultaneously included in the HPDEs. This criterion finds the most parsimonious structure in Λ, i.e. the structure with the largest number of zero elements overall. The orthogonal transformation matrix maximizing the criterion is found by a nonlinear global optimization.

When this technique is applied, it turns out that some of the factors have only few nonzero loadings left. Moreover, it may be possible to find factors with exclusively zero loadings. This implies that the factor is purely spurious, and the number of factors assumed in the model specification should be accordingly reduced. According to the three-indicator rule, see e.g. Anderson and Rubin (1956) or Bollen (1989), a factor cannot be identified unless it has at least three nonzero loadings. Thus if an orthogonal transformation exists that results in HPDEs for all but at most two loadings in the same column containing the zero, there is at least one spurious factor. In Section 5.4.2, the two model selection criteria favored a model with $K = 7$ factors and $K = 1$ factor, respectively. I chose the model with $K = 7$ factors initially, but will now investigate whether any of these factors are in fact spurious. Chapter 4 suggests a criterion that assigns double weights to elements of Λ that are located in a column with less than three nonzero entries, i.e. where all except for two or fewer HPDEs contain the zero simultaneously. Applying this criterion first, I try to determine which factors are spurious in the first step, determine which variables have only nonzero loadings on all non-spurious factors and should therefore be excluded from the analysis, and afterwards find a sparse representation for a model with the accordingly reduced number of factors for the relevant variables.[17]

[17] Note that to identify irrelevant variables, no orthogonal transformation of the HPDEs is necessary, as an HPDE containing the zero simultaneously for the loadings on all factors will not lose this property as the result of an orthogonal transformation, and, conversely, an HPDE not containing the zero simultaneously for the loadings on all factors will not obtain this property as the result of an orthogonal transformation, compare Section 4.4.2 in Chapter 4.

I use three different values of α, namely 0.1, 0.05 and 0.01. These three values of α may result in different numbers of factors identified as spurious and a different sparse structure, where large values of α can be expected to generate less sparsity. Ideally, the results are consistent in a way that the representations with higher degrees of sparsity are nested within those with lower degrees of sparsity, i.e. loadings that have been set to zero for large α should not be different from zero for a small α. Deviations from this may be the result of numerical variation, but may also indicate the presence of multiple sparse patterns, or sparse multimodality. Varying the random number generator's seed, I re-estimate the model with $K = 7$ factors and $P = 1$ lag in the factor process 20 times and determine the number of factors from each estimate. This gives an insight into numerical variation with respect to the number of spurious factors that are identified.

For $\alpha = 0.1$, in 18 out of 20 cases, no column from the loadings matrix contains less than three nonzero elements, hence no factors are found to be spurious, whereas in 2 cases, one spurious factor is found and the number of factors is accordingly reduced to $K = 6$. Thus, for this setting, I choose $K = 7$. For $\alpha = 0.05$, in 20 out of 20 cases, exactly one column of the loadings matrix contains less than three nonzero elements, hence I choose $K = 6$. Eventually, for $\alpha = 0.01$, in 14 out of 20 cases, 3 columns are found to contain less than three nonzero elements, i.e. three factors are found to be spurious, whereas in the remaining 6 cases, only 2 factors are found to be spurious. Hence, I choose $K = 4$ here. As α decreases and the width of the HPDEs $1 - \alpha$ accordingly increases, the number of retained factors decreases, but the degree of sparsity likewise decreases. The remaining non-spurious factors therefore have more nonzero loadings. The model with $\alpha = 0.1$ has a degree of sparsity of 81.38% for its seven factors due to its 524 nonzero loadings out of 2814 total elements in Λ, the model with $\alpha = 0.05$ has a degree of sparsity of 80.02% for its six factors due to its 482 nonzero loadings out of 2412 total elements in Λ, and the model with $\alpha = 0.01$ has a degree of sparsity of 72.82% for its four factors due to its 437 nonzero loadings out of 1608 total elements in Λ. The factors that are retained when α is increased are thus relatively more informative, which may be considered as an argument in favor of the sparse estimation approach. With regard to elimination of irrelevant variables, 25, 31 and 35 counties have only zero loadings for $\alpha = 0.1$, $\alpha = 0.05$ and $\alpha = 0.01$, respectively.

5.6.3 Economic Interpretation of the Sparse Model Estimates

Table 5.10 reports the sparse loadings structures for the three choices of α, where the diagonal elements denote the number of nonzero loadings on each factor, and the elements below the diagonal denote the number of pairwise nonzero loadings, so e.g. for the model with $\alpha = 0.1$, the first factor has 357 nonzero loadings, and there are 33 counties with nonzero loadings on the first and second factor. The first factor in all three sparse models has nonzero loadings for the majority of the counties, so it may be interpreted as a *business cycle factor*. This factor is very similar throughout the different choices for α, and its correlation with GDP growth is -0.6985, -0.6870 and -0.6928, respectively. A comparison to the contemporaneous ifo

business climate indicator yields correlation coefficients of -0.6876, -0.6601 and -0.6726, respectively. The ifo business expectations indicator lagged by two months yields the largest correlation coefficients in magnitude, reaching -0.7458, -0.7184 and -0.7288, respectively. In the model with $\alpha = 0.1$, the fifth, sixth and seventh factor all have nonzero loadings only for counties which also load on the first factor, so if these factors can be interpreted as local business cycles, the cycles occur in addition to the overall cycle.[18] Conversely, the second and third factor have nonzero loadings for some counties that have zero loadings on the first factor, which indicates that if they are interpreted as local business cycles, these cycles partially replace the overall business cycle. In the sparse models with $\alpha = 0.05$ and $\alpha = 0.01$, the second factor has several nonzero loadings that do not coincide with nonzero loadings on the first factor, whereas the other factors have at most one nonzero loading not coinciding with nonzero loadings on the first factor. Further, it must be noted that joint nonzero loadings on more than one of the "local factors" are very rare, so nonzero loadings on the second to last factors are generally mutually exclusive for each variable.

The factors and corresponding loadings from the models with different values for α have been ordered such that the models with larger α - and hence a higher number of factors - nest the models with smaller α - and hence a smaller number of factors. Comparing the factors, which are shown for the three models together with their corresponding HPDIs and the business climate indicator for comparison with the first factor in Figures 5.14, 5.16 and 5.18, reveals that the first three factors are very similar throughout, see Table 5.11. The correlation of the business cycle factor from each of the models with the corresponding factor from the other models always exceeds 0.997, the correlation of the second and third factor from each model with their counterparts from the other models lies between 0.89 and 0.99. These three factors are also very similar to the first, fifth and third factor from the sparse model estimated at the beginning of this section which used univariate HPDIs for the loadings parameters corresponding to the factors having maximal kurtosis. The pairwise correlations lie between 0.85 and 0.95, and the loadings structure also looks similar. The respective pairwise correlation of the remaining factors from the three sparse models are substantially smaller, lying between 0.57 and 0.83.

Accordingly, the HPDIs for the first three factors are substantially narrower than those for the remaining factors, see Table 5.12. The average width of the 68% HPDI for the business cycle factor is smaller than 0.25 for all three models, the average widths of the corresponding HPDIs for the second and third factor are between 0.6 and 0.8. The average widths of the HPDIs for the remaining factors all exceed 1 and are sometimes even larger than 2. Figures 5.15, 5.17 and 5.19 show the sparse loadings patterns, which resemble each other, particularly for the first three factors. As shown in Table 5.10, several counties with zero loadings on the business cycle factor have nonzero loadings on the second and third factor. The aforementioned division between the South and the West on the one side, and the North and the East on the other side is clearly visible here. While the counties in the extreme eastern and north eastern parts of the country often have zero loadings on the business cycle factor, hence replacing it

[18] As for the fourth factor, only one variable has a nonzero loading that has a zero loading for the first factor.

with a local factor, a broad strip located further towards the Southwest has generally lower loadings on the business cycle factor and additional nonzero loadings on the second local factor. These factors have overall nonzero loadings with the same signs, indicating that they indeed represent secondary cycles. As for the remaining factors, there are only few nonzero loadings, and the number of negative nonzero loadings is similar to that of positive nonzero loadings. Thus these factors appear to combine multiple effects that are locally constrained to very small regions. A look at the factors themselves shows that their distinct features are mostly a few spikes, similar to those found for the factors that were rotated to maximize the kurtosis in Section 5.5.

Table 5.9 reports the average loadings for the sparse model with $\alpha = 0.05$ distinguished by country (Bundesland). The results look overall similar to those from Table 5.6, but are even more pronounced. The first factor, which can be understood as the business cycle factor, plays an important role for almost all countries (Bundesländer), except for Brandenburg, Mecklenburg-Vorpommern and Saxony-Anhalt, and, to a lesser extent, for Schleswig-Holstein. The second factor, which can be understood as the first "local factor", is unimportant for most countries (Bundesländer), except for Brandenburg and Mecklenburg-Vorpommern. The third factor, which can be understood as the second "local factor" is likewise unimportant for most countries (Bundesländer), except for Lower Saxony and Bremen. The remaining three factors show no distinctive patterns for any countries (Bundesländer). The fact that the local factors are somewhat related to the country boundaries indicates that they might, at least in part, be the result of local policies. It is not immediately clear, however, which policies could underly these patterns.

Eventually, Figure 5.20 shows the explanatory power of the three models with different choices of α, i.e. the share of variation per county explained by the factors. When compared to the results for the full model from Section 5.5, the explanatory power is smaller on average, but for the model with $\alpha = 0.1$, it is larger for 61 counties, for the model with $\alpha = 0.05$, it is larger for 40 counties, and for the model with $\alpha = 0.01$, it is larger for 33 counties.

5.7 Forecasting Exercise

In this section, I briefly evaluate the forecasting performance of the full and the sparse factor model by comparing it to the forecasting performance of a simple AR(1) model for the individual observable series y_i with $i \in \{1, \dots, 402\}$. The forecasting experiment is conducted as follows: A growing sample containing the first $\mathfrak{t}$ observations of each series with $\mathfrak{t} \in \{50, \dots, 78\}$ is used to sample from the forecast densities of $y_{i,\mathfrak{t}+\tau}$ for all $i \in \{1, \dots, 402\}$ and for forecast horizons $\tau \in \{1, \dots, 4\}$. To distinguish them from the observed data, the forecasts are denoted as $\hat{y}_{i,\mathfrak{t}+\tau}$. The according samples are directly obtained from the Gibbs sampler, which is slightly extended for this purpose. The samples each have size Z, which is identical to the length of the retained Gibbs sequence obtained from the sampler, chosen as $Z = 10,000$, compare Section 5.3.

The means of the generated samples $\bar{\hat{y}}_{i,\mathfrak{t}+\tau} = \frac{1}{Z}\sum_{z=1}^{Z} \hat{y}_{i,\mathfrak{t}+\tau}^{(z)}$ for the full and the sparse factor model and the AR(1) model for each observed series are then compared to the realized $y_{i,\mathfrak{t}+\tau}$ by calculating the root mean-squared forecasting error (RMSFE) for the forecast horizon τ, i.e.

$$RMSFE_{i,\tau} = \sqrt{\frac{1}{29}\sum_{\mathfrak{t}=50}^{78}(\bar{\hat{y}}_{i,\mathfrak{t}+\tau} - y_{i,\mathfrak{t}+\tau})^2}, \tag{5.69}$$

which yields one RMSFE for each of the 402 series at each of the 4 forecast horizons.

The full factor model is estimated for each of the growing subsamples with $K = 7$, $P = 1$ and $Q = 0$, as chosen by the marginal likelihood model selection criterion. After convergence has been detected, in each iteration z, the sampled $\Phi^{(z)}$ and the sampled factors $\{f_t^{(z)}\}_{t=1}^{\mathfrak{t}}$ are used to calculate forecasts for the factors as $\hat{f}_{\mathfrak{t}+\tau}^{(z)} = \Phi^{(z)} f_{\mathfrak{t}}^{(z)}$. Using the sampled $\Lambda^{(z)}$ and $\sigma_i^{2(z)}$, the forecast for $y_{i,\mathfrak{t}+\tau}$ is then obtained as $\hat{y}_{i,\mathfrak{t}+\tau}^{(z)} = \Lambda^{(z)}\hat{f}_{\mathfrak{t}+\tau}^{(z)} + \tilde{\xi}_{i,\mathfrak{t}+\tau}^{(z)}$, where the auxiliary idiosyncratic component $\tilde{\xi}_{i,\mathfrak{t}+\tau}^{(z)}$ is sampled from $N(0, \sigma_i^{2(z)})$. Note that the full factor model does not require any identification constraints, as the factors and loadings are not considered separately from each other and the rotation problem therefore does not occur.

The sparse factor model likewise starts with $K = 7$, $P = 1$ and $Q = 0$ and initially determines the sparse loadings structure as described in Section 5.6. As the sparse structure is determined individually for each of the growing subsamples, it changes as the size of the available sample increases. Using $\alpha = 0.05$ throughout, the number of factors varies between 5 and 6, whereas the loadings structure generally looks similar, in particular for the three factors with the largest number of nonzero loadings. In the subsequent confirmatory factor analysis, a sample from the $\hat{y}_{i,\mathfrak{t}+\tau}^{(z)}$ is obtained in the same fashion as for the full factor model.

The sampling approach for the AR(1) models is straightforward and follows Koop (2003). For the persistence parameters per series ρ_i with $i \in \{1, \ldots, N\}$, the prior distribution is a normal distribution with prior hyperparameters μ_{ρ_i} and $\sigma_{\rho_i}^2$,[19] and for the variance parameters per series ψ_i^2, the prior distribution is an inverse Gamma distribution with prior hyperparameters $\underline{\alpha}_{\psi_i}$ and $\underline{\beta}_{\psi_i}$. The ρ_i and ψ_i^2 are alternatingly sampled from their full conditional posterior distributions, and the ψ_i^2 are used to sample auxiliary idiosyncratic terms that are added to the autoregression result to obtain $\hat{y}_{i,\mathfrak{t}+\tau}^{(z)}$ from the forecast density in each iteration z after convergence of the Gibbs sampler. The prior hyperparameters are chosen as $\mu_{\rho_i} = 0$, $\sigma_{\rho_i}^2 = 1$, $\underline{\alpha}_{\psi_i} = 1$ and $\underline{\beta}_{\psi_i} = 1$ for all $i \in \{1, \ldots, N\}$.

The AR(1) approach serves as the reference for the two factor model approaches, whose RMSFEs are shown relative to the reference RMSFEs in Figure 5.21. The RMSFEs for the full factor model are about 50% larger on average for $\tau = 1$, about 80% larger on average for $\tau = 2$, three times larger for $\tau = 3$ and twelve times larger for $\tau = 4$. Whereas the results are similar for the sparse model for $\tau = 1$, the RMSFEs for $\tau = 2$ and $\tau = 3$ are both also only

[19] To ensure stationarity, a truncated normal prior should be used. The growth rates analyzed here, however, are nowhere near nonstationarity, compare Section 5.4.1, so the truncation has virtually no effect. The truncated normal prior can easily be implemented by rejecting draws where $|\rho_i| > 0.9999$.

about 50% larger than those from the reference model on average. For $\tau = 4$, the RMSFEs are about two and a half times larger on average. Hence the forecasting performance of the factor models in this application does not beat the AR(1) model, however, the sparse factor model performs substantially better than the full model.

5.8 Conclusion

In this chapter, I discuss an estimation approach for dynamic factor models with autocorrelated idiosyncratic components which uses on the WOP ex-post identification scheme from Chapter 3. The approach is applied to a data set of unemployment figures for 402 German counties, observed over 82 periods. The data are transformed to growth rates, which results in their stationarity. Next, I compare different model specifications using the DIC and the marginal likelihood determined by Chib's method, see Chib (1995). The marginal likelihood indicates that a model with $K = 7$ factors and $P = 1$ lag in the factor structure is suitable for the data set, whereas the DIC favors a model with $K = 1$ and $P = 1$. Later on, a method to identify spurious factors is applied, so I opt for the model with $K = 7$. Estimating this model, I find that the factors can be transformed such that they highly correlate with the first seven principal components. The first factor can be rotated to have a high negative correlation with several business-cycle related figures, particularly with the ifo business climate indicator. The remaining factors are harder to interpret. Thus I run the approach from Chapter 4 to identify spurious factors and irrelevant variables and to find a sparse pattern in the loadings matrix. For three different choices of α, which determines the width of the multivariate HPDIs based on which elements of the loadings matrix are set to zero, up to three factors are found to be spurious. The first three non-spurious factors in the sparse model resemble each other for all three different values of α, the first having nonzero loadings for the majority of the counties, thus interpreted as a business cycle factor, whereas the second and third show a local pattern. Where the first of these local factors has a nonzero loading, it appears to replace the business cycle factor, whereas the second local factor complements the business cycle factor. As the nonzero loadings on the two local factors occur particularly in certain countries (Bundesländer), they may partially be induced by regional policies.

In a short forecasting exercise, I find that neither the full nor the sparse factor model succeeds at outperforming a simple AR(1) model, which is used as a reference. The RMSFEs are substantially larger for the factor models, however, the sparse factor model performs substantially better than the full factor model, especially at increased forecast horizons.

Tables

1	Flensburg	135	Marburg-Biedenkopf	269	Bamberg
2	Kiel	136	Vogelsbergkreis	270	Bayreuth
3	Lübeck	137	Kassel	271	Coburg
4	Neumünster	138	Fulda	272	Hof
5	Dithmarschen	139	Hersfeld-Rotenburg	273	Bamberg
6	Herzogtum Lauenburg	140	Kassel	274	Bayreuth
7	Nordfriesland	141	Schwalm-Eder-Kreis	275	Coburg
8	Ostholstein	142	Waldeck-Frankenberg	276	Forchheim
9	Plön	143	Werra-Meißner-Kreis	277	Hof
10	Rendsburg-Eckernförde	144	Koblenz	278	Kronach
11	Schleswig-Flensburg	145	Ahrweiler	279	Kulmbach
12	Segeberg	146	Altenkirchen (Westerw.)	280	Lichtenfels
13	Steinburg	147	Bad Kreuznach	281	Wunsiedel i. Fichtelgeb.
14	Stormarn	148	Birkenfeld	282	Ansbach
15	Hamburg	149	Cochem-Zell	283	Erlangen
16	Braunschweig	150	Mayen-Koblenz	284	Fürth
17	Salzgitter	151	Neuwied	285	Nürnberg
18	Wolfsburg	152	Rhein-Hunsrück-Kreis	286	Schwabach
19	Gifhorn	153	Rhein-Lahn-Kreis	287	Ansbach
20	Göttingen	154	Westerwaldkreis	288	Erlangen-Höchstadt
21	Goslar	155	Trier	289	Fürth
22	Helmstedt	156	Bernkastel-Wittlich	290	Nürnberger Land
23	Northeim	157	Eifelkreis-Bitburg-Prüm	291	Neustadt a.d. Aisch-Bad Windsh.
24	Osterode am Harz	158	Vulkaneifel	292	Roth
25	Peine	159	Trier-Saarburg	293	Weißenburg-Gunzenhausen
26	Wolfenbüttel	160	Frankenthal (Pfalz)	294	Aschaffenburg
27	Region Hannover	161	Kaiserslautern	295	Schweinfurt
28	Diepholz	162	Landau in der Pfalz	296	Würzburg
29	Hameln-Pyrmont	163	Ludwigshafen am Rhein	297	Aschaffenburg
30	Hildesheim	164	Mainz	298	Bad Kissingen
31	Holzminden	165	Neustadt a.d. Weinstraße	299	Rhön-Grabfeld
32	Pinneberg	166	Pirmasens	300	Haßberge
33	Nienburg (Weser)	167	Speyer	301	Kitzingen
34	Schaumburg	168	Worms	302	Miltenberg
35	Celle	169	Zweibrücken	303	Main-Spessart
36	Cuxhaven	170	Alzey-Worms	304	Schweinfurt
37	Harburg	171	Bad Dürkheim	305	Würzburg
38	Lüchow-Dannenberg	172	Donnersbergkreis	306	Augsburg
39	Lüneburg	173	Germersheim	307	Kaufbeuren
40	Osterholz	174	Kaiserslautern	308	Kempten (Allgäu)
41	Rotenburg (Wümme)	175	Kusel	309	Memmingen
42	Heidekreis	176	Südliche Weinstraße	310	Aichach-Friedberg
43	Stade	177	Rhein-Pfalz-Kreis	311	Augsburg
44	Uelzen	178	Mainz-Bingen	312	Dillingen a.d. Donau
45	Verden	179	Südwestpfalz	313	Günzburg
46	Delmenhorst	180	Stuttgart	314	Neu-Ulm
47	Emden	181	Böblingen	315	Lindau (Bodensee)
48	Oldenburg (Oldenburg)	182	Esslingen	316	Ostallgäu
49	Osnabrück	183	Göppingen	317	Unterallgäu
50	Wilhelmshaven	184	Ludwigsburg	318	Donau-Ries
51	Ammerland	185	Rems-Murr-Kreis	319	Oberallgäu
52	Aurich	186	Heilbronn	320	Stadtverband Saarbrücken
53	Cloppenburg	187	Heilbronn	321	Merzig-Wadern
54	Emsland	188	Hohenlohekreis	322	Neunkirchen
55	Friesland	189	Schwäbisch Hall	323	Saarlouis
56	Grafschaft Bentheim	190	Main-Tauber-Kreis	324	Saarpfalz-Kreis
57	Leer	191	Heidenheim	325	St. Wendel
58	Oldenburg	192	Ostalbkreis	326	Berlin
59	Osnabrück	193	Baden-Baden	327	Brandenburg an der Havel
60	Vechta	194	Karlsruhe	328	Cottbus
61	Wesermarsch	195	Karlsruhe	329	Frankfurt (Oder)
62	Wittmund	196	Rastatt	330	Potsdam

63	Bremen	197	Heidelberg	331	Barnim
64	Bremerhaven	198	Mannheim	332	Dahme-Spreewald
65	Düsseldorf	199	Neckar-Odenwald-Kreis	333	Elbe-Elster
66	Duisburg	200	Rhein-Neckar-Kreis	334	Havelland
67	Essen	201	Pforzheim	335	Märkisch-Oderland
68	Krefeld	202	Calw	336	Oberhavel
69	Mönchengladbach	203	Enzkreis	337	Oberspreewald-Lausitz
70	Mülheim an der Ruhr	204	Freudenstadt	338	Oder-Spree
71	Oberhausen	205	Freiburg im Breisgau	339	Ostprignitz-Ruppin
72	Remscheid	206	Breisgau-Hochschwarzw.	340	Potsdam-Mittelmark
73	Solingen	207	Emmendingen	341	Prignitz
74	Wuppertal	208	Ortenaukreis	342	Spree-Neiße
75	Kleve	209	Rottweil	343	Teltow-Fläming
76	Mettmann	210	Schwarzwald-Baar-Kreis	344	Uckermark
77	Rhein-Kreis Neuss	211	Tuttlingen	345	Rostock
78	Viersen	212	Konstanz	346	Schwerin
79	Wesel	213	Lörrach	347	Rostock
80	Bonn	214	Waldshut	348	Ludwigslust-Parchim
81	Köln	215	Reutlingen	349	Mecklenburgische Seenplatte
82	Leverkusen	216	Tübingen	350	Vorpommern-Rügen
83	Aachen	217	Zollernalbkreis	351	Nordwestmecklenburg
84	Düren	218	Ulm	352	Vorpommern-Greifswald
85	Rhein-Erft-Kreis	219	Alb-Donau-Kreis	353	Chemnitz
86	Euskirchen	220	Biberach	354	Mittelsachsen
87	Heinsberg	221	Bodenseekreis	355	Vogtlandkreis
88	Oberbergischer Kreis	222	Ravensburg	356	Erzgebirgskreis
89	Rheinisch-Bergischer Kr.	223	Sigmaringen	357	Zwickau
90	Rhein-Sieg-Kreis	224	Ingolstadt	358	Dresden
91	Bottrop	225	München	359	Görlitz
92	Gelsenkirchen	226	Rosenheim	360	Bautzen
93	Münster	227	Altötting	361	Meißen
94	Borken	228	Berchtesgadener Land	362	Sächsische Schweiz-Osterzgebirge
95	Coesfeld	229	Bad Tölz-Wolfratshausen	363	Leipzig
96	Recklinghausen	230	Dachau	364	Leipzig
97	Steinfurt	231	Ebersberg	365	Nordsachsen
98	Warendorf	232	Eichstätt	366	Dessau-Roßlau
99	Bielefeld	233	Erding	367	Anhalt-Bitterfeld
100	Gütersloh	234	Freising	368	Wittenberg
101	Herford	235	Fürstenfeldbruck	369	Halle (Saale)
102	Höxter	236	Garmisch-Partenkirchen	370	Burgenlandkreis
103	Lippe	237	Landsberg am Lech	371	Saalekreis
104	Minden-Lübbecke	238	Miesbach	372	Mansfeld-Südharz
105	Paderborn	239	Mühldorf a. Inn	373	Magdeburg
106	Bochum	240	München	374	Harz
107	Dortmund	241	Neuburg-Schrobenhausen	375	Jerichower Land
108	Hagen	242	Pfaffenhofen a.d. Ilm	376	Börde
109	Hamm	243	Rosenheim	377	Stendal
110	Herne	244	Starnberg	378	Salzlandkreis
111	Ennepe-Ruhr-Kreis	245	Traunstein	379	Altmarkkreis Salzwedel
112	Hochsauerlandkreis	246	Weilheim-Schongau	380	Erfurt
113	Märkischer Kreis	247	Landshut	381	Gera
114	Olpe	248	Passau	382	Jena
115	Siegen-Wittgenstein	249	Straubing	383	Suhl
116	Soest	250	Deggendorf	384	Weimar
117	Unna	251	Freyung-Grafenau	385	Eisenach
118	Darmstadt	252	Kelheim	386	Eichsfeld
119	Frankfurt am Main	253	Landshut	387	Nordhausen
120	Offenbach am Main	254	Passau	388	Wartburgkreis
121	Wiesbaden	255	Regen	389	Unstrut-Hainich-Kreis
122	Bergstraße	256	Rottal-Inn	390	Kyffhäuserkreis
123	Darmstadt-Dieburg	257	Straubing-Bogen	391	Schmalkalden-Meiningen
124	Groß-Gerau	258	Dingolfing-Landau	392	Gotha
125	Hochtaunuskreis	259	Amberg	393	Sömmerda
126	Main-Kinzig-Kreis	260	Regensburg	394	Hildburghausen
127	Main-Taunus-Kreis	261	Weiden i.d. OPf.	395	Ilm-Kreis

128	Odenwaldkreis	262	Amberg-Sulzbach	396	Weimarer Land
129	Offenbach	263	Cham	397	Sonneberg
130	Rheingau-Taunus-Kreis	264	Neumarkt i.d. OPf.	398	Saalfeld-Rudolstadt
131	Wetteraukreis	265	Neustadt a.d. Waldnaab	399	Saale-Holzland-Kreis
132	Gießen	266	Regensburg	400	Saale-Orla-Kreis
133	Lahn-Dill-Kreis	267	Schwandorf	401	Greiz
134	Limburg-Weilburg	268	Tirschenreuth	402	Altenburger Land

Table 5.1: Names of the 402 counties or NUTS-3 regions of Germany contained in the sample.

P	0	1	2	3	4
K			$Q = 0$		
1	94120.98	**94108.57**	94121.45	94130.80	94138.64
2	94810.25	94810.00	94828.60	94885.56	94936.39
3	95503.14	95524.39	95577.14	95672.48	95790.56
4	96191.18	96234.82	96344.76	96461.26	96627.17
5	96892.27	96975.70	97128.91	97319.12	97529.26
6	97592.97	97698.58	97875.84	98123.15	98407.23
7	98278.91	98419.16	98672.91	98951.56	99313.46
P	0	1	2	3	4
K			$Q = 1$		
1	94268.78	94289.48	94296.36	94300.72	94311.85
2	95069.28	95098.74	95128.28	95163.47	95217.09
3	95717.95	95754.44	95796.48	95889.13	95979.77
4	96441.39	96501.65	96611.91	96736.48	96896.96
5	97123.75	97211.93	97360.91	97537.48	97752.60
6	97811.58	97925.26	98113.11	98349.68	98622.95
7	98489.72	98628.46	98863.09	99140.67	99485.83
P	0	1	2	3	4
K			$Q = 2$		
1	94679.68	94717.80	94716.42	94726.45	94741.63
2	95503.35	95588.22	95649.55	95751.16	95840.72
3	96126.36	96201.17	96266.45	96412.74	96559.21
4	96779.31	96887.69	97114.46	97297.36	97550.10
5	97469.90	97620.78	97894.78	98142.25	98427.02
6	98134.91	98311.10	98624.20	98930.92	99286.63
7	98806.20	98988.70	99317.50	99672.53	100087.58

Table 5.2: Deviance Information Criterion (DIC) for different models.

P	0	1	2	3	4
K			$Q = 0$		
1	11577.96	11600.07	11600.45	11598.94	11604.30
2	11930.95	11948.17	11942.13	11934.64	11940.78
3	12071.31	12060.88	12069.77	12050.00	12023.03
4	12440.96	12416.24	12424.66	12415.36	12398.54
5	12556.36	12517.70	12503.57	12479.64	12477.77
6	12754.44	12706.47	12698.72	12659.88	12630.34
7	**12936.23**	12903.88	12871.39	12862.22	12847.28
P	0	1	2	3	4
K			$Q = 1$		
1	10985.55	11058.53	11063.00	11068.97	11072.57
2	11422.64	11512.97	11511.17	11515.53	11503.45
3	11518.17	11605.53	11612.31	11607.27	11589.76
4	11863.10	11951.27	11939.12	11923.96	11909.25
5	11972.98	12047.66	12044.30	12043.54	12024.75
6	12194.73	12244.89	12235.37	12200.89	12197.14
7	12346.14	12390.01	12373.52	12351.99	12347.44
P	0	1	2	3	4
K			$Q = 2$		
1	10806.47	10889.38	10897.16	10889.59	10895.35
2	11211.15	11249.45	11268.53	11226.28	11208.92
3	11274.85	11348.88	10858.51	11000.22	11248.33
4	10990.16	11517.81	10894.88	11295.53	11513.94
5	10870.97	11337.34	11514.10	11259.38	11032.29
6	11429.37	11552.99	11616.94	11750.62	11770.20
7	11309.92	11877.37	11636.26	11938.37	11877.86

Table 5.3: Log marginal likelihood $\log(p(Y|m))$ for different models.

Factor	1	2	3	4	5	6	7
Kurtosis	3.6975	28.7627	11.3705	10.3779	9.8206	6.2966	2.4561
Explained variation	0.1823	0.0145	0.0258	0.1156	0.0289	0.0156	0.0261
	(0.1551)	(0.0213)	(0.0430)	(0.0678)	(0.0364)	(0.0208)	(0.0298)
Most relevant factor	221	1	27	129	17	2	5

Table 5.4: Kurtosis, explained variation, and number of variables for which the rotated factor is the most relevant one, explaining the largest share of the total explained variance. Standard deviations in parentheses.

Factor	1	2	3	4	5	6	7
Positive loadings	385	307	247	401	365	254	363

Table 5.5: Number of positive entries in the loadings matrix per factor.

Notes: The remaining loadings are negative.

Factor	1	2	3	4	5	6	7
Schleswig-Holstein	0.1764	0.0001	−0.0888	0.3269	0.2530	0.0513	0.1207
	(0.1573)	(0.0792)	(0.0805)	(0.1014)	(0.0851)	(0.0560)	(0.0604)
Hamburg	0.4556	0.0295	−0.1182	0.3007	0.3154	−0.0250	0.1271
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Lower Saxony	0.2678	0.0086	0.3102	0.3273	0.1677	0.0981	0.0177
	(0.1133)	(0.0830)	(0.1038)	(0.1075)	(0.0910)	(0.0751)	(0.0616)
Bremen	0.2707	0.0081	0.2255	0.3493	0.1157	−0.0624	0.0126
	(0.2731)	(0.0154)	(0.0591)	(0.0495)	(0.0837)	(0.0550)	(0.0175)
North-Rhine Westphalia	0.4820	0.0577	−0.0275	0.3225	0.1301	−0.0130	0.0217
	(0.1600)	(0.0887)	(0.0600)	(0.0912)	(0.0920)	(0.0558)	(0.0646)
Hesse	0.3128	0.0660	0.0011	0.3150	0.1191	−0.1731	0.1224
	(0.1291)	(0.0929)	(0.0876)	(0.1247)	(0.0958)	(0.0914)	(0.0797)
Rhineland Palatinate	0.3318	0.0486	0.0163	0.3072	0.1364	−0.1837	0.1765
	(0.1129)	(0.0619)	(0.0840)	(0.0728)	(0.0894)	(0.0847)	(0.0630)
Baden-Württemberg	0.6228	0.0786	−0.0040	0.2897	0.0764	0.0369	0.0995
	(0.1232)	(0.0916)	(0.0721)	(0.0988)	(0.0857)	(0.0626)	(0.0549)
Bavaria	0.4865	0.1376	0.0139	0.3439	0.0678	0.1281	0.1988
	(0.1196)	(0.0815)	(0.0716)	(0.1157)	(0.0760)	(0.0724)	(0.0901)
Saarland	0.4084	0.0603	0.0401	0.3300	0.0671	−0.1054	0.1379
	(0.1178)	(0.1958)	(0.0913)	(0.0999)	(0.0262)	(0.0580)	(0.0745)
Berlin	0.3069	−0.0173	−0.1146	0.4745	0.1452	0.1325	0.0676
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Brandenburg	0.0271	−0.0314	−0.0557	0.3491	0.2512	0.0368	0.1744
	(0.1261)	(0.1073)	(0.0988)	(0.0828)	(0.1211)	(0.0603)	(0.0873)
Mecklenburg-Vorpommern	0.0291	−0.0245	−0.0592	0.4297	0.2470	−0.0157	0.1481
	(0.0499)	(0.0811)	(0.0535)	(0.1013)	(0.1233)	(0.0720)	(0.0907)
Saxony	0.2610	0.0153	0.2593	0.3174	0.1389	0.0122	0.2020
	(0.1039)	(0.1036)	(0.0738)	(0.0826)	(0.0935)	(0.0430)	(0.0807)
Saxony-Anhalt	0.1429	0.0391	0.2193	0.2494	0.1812	0.0491	0.1423
	(0.0910)	(0.1108)	(0.0831)	(0.1051)	(0.0946)	(0.0748)	(0.0615)
Thuringia	0.2727	0.0666	0.2400	0.3110	0.2054	0.0095	0.1714
	(0.1446)	(0.0897)	(0.0879)	(0.0895)	(0.1104)	(0.0747)	(0.0620)
Germany	0.3755	0.0634	0.0612	0.3234	0.1315	0.0209	0.1260
	(0.2033)	(0.1022)	(0.1484)	(0.1050)	(0.1075)	(0.1232)	(0.1009)

Table 5.6: Average loadings per country (Bundesland) and for all of Germany for the rotated factors. Standard deviations in parentheses.

Factor	1	2	3	4	5	6	7
Zero loadings (68% HPDI)	35	184	200	6	111	179	119
Zero loadings (90% HPDI)	56	269	267	17	193	269	192
Zero loadings (95% HPDI)	71	300	292	34	230	295	225
Positive loadings	330	90	96	269	170	62	175
Negative loadings	1	12	14	99	2	45	2
Sign different	0	0	0	99	1	0	1

Table 5.7: Zero loadings identified by the HPDIs for the rotated factors, positive and negative loadings in the estimated sparse model (based on the 95% HPDI), and number of cases where the sign in the sparse model is different from that in the full model.

Factor	1	2	3	4	5	6	7
68% HPDI	1.5629	1.2164	1.3117	1.0491	1.5057	1.1045	1.6705
	(0.0761)	(0.0974)	(0.0797)	(0.0365)	(0.1131)	(0.0639)	(0.1093)
90% HPDI	1.5561	1.2141	1.3097	1.0485	1.4995	1.0995	1.6686
	(0.0768)	(0.0969)	(0.0818)	(0.0362)	(0.1115)	(0.0637)	(0.1106)
95% HPDI	1.5513	1.2110	1.3076	1.0469	1.4951	1.0993	1.6609
	(0.0789)	(0.0984)	(0.0820)	(0.0363)	(0.1112)	(0.0638)	(0.1183)

Table 5.8: Ratio between the HPDI widths for the factors from the full and the sparse model.

Factor	1	2	3	4	5	6	7
Schleswig-Holstein	0.1742	−0.0175	−0.0523	0.0977	0.3278	0.0000	0.0679
	(0.1782)	(0.0631)	(0.1012)	(0.1130)	(0.1513)	(0.0000)	(0.1105)
Hamburg	0.5129	0.0000	0.0000	−0.0485	0.3732	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Lower Saxony	0.1647	0.0152	0.4134	0.0947	0.1013	0.0437	0.0027
	(0.1307)	(0.0856)	(0.1249)	(0.1028)	(0.1052)	(0.0879)	(0.0181)
Bremen	0.2537	0.0000	0.1889	0.1523	0.0511	0.0000	0.0000
	(0.2537)	(0.0000)	(0.1889)	(0.2067)	(0.0511)	(0.0000)	(0.0000)
North-Rhine Westphalia	0.4874	0.0114	−0.0071	0.1359	0.0861	−0.0071	−0.0036
	(0.1931)	(0.0995)	(0.0361)	(0.1209)	(0.1127)	(0.0360)	(0.0457)
Hesse	0.2672	0.0352	−0.0017	0.1169	0.0938	−0.1961	0.0587
	(0.1626)	(0.1050)	(0.0570)	(0.1239)	(0.1156)	(0.1677)	(0.1074)
Rhineland Palatinate	0.2829	0.0064	0.0021	0.0430	0.0988	−0.2340	0.1458
	(0.1500)	(0.0384)	(0.0535)	(0.0966)	(0.1251)	(0.1594)	(0.1176)
Baden-Württemberg	0.6491	0.0352	−0.0115	0.0458	0.0321	0.0143	0.0451
	(0.1415)	(0.1144)	(0.0427)	(0.1125)	(0.0806)	(0.0404)	(0.0764)
Bavaria	0.4311	0.0888	0.0067	0.0684	0.0308	0.0800	0.2420
	(0.1620)	(0.1066)	(0.0526)	(0.1320)	(0.0673)	(0.0917)	(0.1636)
Saarland	0.3544	0.0125	0.0397	0.1090	0.0000	−0.0487	0.1474
	(0.1664)	(0.2691)	(0.0889)	(0.1063)	(0.0000)	(0.1089)	(0.1214)
Berlin	0.2871	0.0000	0.0000	0.3455	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Brandenburg	0.0302	−0.0084	−0.0264	0.0734	0.3552	0.0128	0.1623
	(0.1292)	(0.1175)	(0.0748)	(0.1121)	(0.2026)	(0.0527)	(0.1553)
Mecklenburg-Vorpommern	0.0000	−0.0290	0.0000	0.1910	0.3425	0.0000	0.1099
	(0.0000)	(0.0766)	(0.0000)	(0.1334)	(0.1737)	(0.0000)	(0.1494)
Saxony	0.1420	0.0434	0.3226	−0.0185	0.0878	0.0000	0.2145
	(0.1100)	(0.1018)	(0.1086)	(0.1100)	(0.1363)	(0.0000)	(0.1604)
Saxony-Anhalt	0.0388	−0.0175	0.2435	0.0013	0.1671	0.0131	0.1095
	(0.0987)	(0.1249)	(0.1625)	(0.1072)	(0.1575)	(0.0474)	(0.1496)
Thuringia	0.2010	0.0363	0.2703	−0.0119	0.1668	−0.0132	0.1414
	(0.1422)	(0.1204)	(0.1492)	(0.0932)	(0.1726)	(0.0618)	(0.1159)
Germany	0.3336	0.0326	0.0807	0.0739	0.1025	−0.0093	0.1125
	(0.2372)	(0.1098)	(0.1739)	(0.1259)	(0.1476)	(0.1309)	(0.1502)
Germany (nonzeros only)	0.4051	0.1285	0.2951	0.0808	0.2396	−0.0351	0.2555
	(0.1986)	(0.1885)	(0.2183)	(0.1296)	(0.1349)	(0.2531)	(0.1217)

Table 5.9: Average loadings per country (Bundesland) and for all of Germany for the sparse factor model using the sparsity structure identified from the rotated factors.

90% HPDI - $K = 7$						
357						
33	48					
67	6	75				
10	3	0	11			
20	3	2	2	20		
4	0	1	0	1	4	
9	1	2	0	2	1	9

95% HPDI - $K = 6$					
356					
49	63				
31	3	32			
15	3	2	15		
11	1	1	0	11	
5	1	1	2	0	5

99% HPDI - $K = 4$			
349			
31	47		
25	3	26	
14	2	2	15

Table 5.10: Sparse loadings structure for different choices of α.

Notes: Diagonal elements indicate the number of nonzero loadings per factor, remaining elements indicate the number of nonzero loadings for each pair of factors.

Factor	1	2	3	4	5	6
$\alpha = 0.1$ and $\alpha = 0.05$	0.9977	0.9333	0.8949	0.8279	0.5922	0.1364
$\alpha = 0.1$ and $\alpha = 0.01$	0.9982	0.9544	0.9316	0.5061		
$\alpha = 0.05$ and $\alpha = 0.01$	0.9996	0.9863	0.9478	0.5730		

Table 5.11: Correlation between the factors from the sparse models with different values of α.

Notes: Factors have been resorted to maximize the pairwise correlations.

Factor	1	2	3	4	5	6	7
$\alpha = 0.10$	0.2457	0.6837	0.6514	1.6466	1.3946	2.0024	1.3084
$\alpha = 0.05$	0.2345	0.6523	0.7784	1.4602	1.5510	2.1096	
$\alpha = 0.01$	0.2341	0.5989	0.7691	1.3224			

Table 5.12: Average width of the 68% HPDIs for the factors.

Factor	1	2	3	4	5	6
Schleswig-Holstein	0.2852	0.1085	0.0000	−0.0227	0.0000	0.0000
	(0.2274)	(0.2088)	(0.0000)	(0.0819)	(0.0000)	(0.0000)
Hamburg	0.5945	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Lower Saxony	0.3702	0.0121	0.2562	−0.0048	−0.0047	0.0000
	(0.1614)	(0.0467)	(0.2452)	(0.0328)	(0.0316)	(0.0000)
Bremen	0.2938	0.0000	0.1583	0.0000	0.0000	0.0000
	(0.2938)	(0.0000)	(0.1583)	(0.0000)	(0.0000)	(0.0000)
North-Rhine Westphalia	0.5759	−0.0198	0.0000	−0.0219	0.0000	0.0000
	(0.1810)	(0.0698)	(0.0000)	(0.0784)	(0.0000)	(0.0000)
Hesse	0.3831	0.0703	−0.0085	−0.0153	0.0309	0.0299
	(0.2084)	(0.1359)	(0.0423)	(0.0765)	(0.1074)	(0.1041)
Rhineland Palatinate	0.4664	0.0098	−0.0053	0.0000	0.0509	0.0092
	(0.1321)	(0.0578)	(0.0311)	(0.0000)	(0.1290)	(0.0542)
Baden-Württemberg	0.7081	−0.0773	0.0000	−0.0040	0.0000	0.0000
	(0.1595)	(0.1288)	(0.0000)	(0.0264)	(0.0000)	(0.0000)
Bavaria	0.6400	−0.0059	−0.0054	0.0148	−0.0040	−0.0053
	(0.1604)	(0.0544)	(0.0308)	(0.0633)	(0.0277)	(0.0361)
Saarland	0.5381	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0982)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Berlin	0.5365	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Brandenburg	0.0980	0.3815	0.0000	−0.0172	0.0000	0.0000
	(0.1441)	(0.2510)	(0.0000)	(0.0710)	(0.0000)	(0.0000)
Mecklenburg-Vorpommern	0.1252	0.4259	0.0000	0.0000	0.0000	0.0000
	(0.1155)	(0.2514)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Saxony	0.3845	0.0597	0.0000	0.0150	0.0000	0.0000
	(0.1950)	(0.1467)	(0.0000)	(0.0541)	(0.0000)	(0.0000)
Saxony-Anhalt	0.2140	0.0906	0.0000	0.0000	0.0000	0.0000
	(0.1807)	(0.1666)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Thuringia	0.4240	0.1217	0.0222	0.0000	0.0100	0.0000
	(0.1632)	(0.2109)	(0.1039)	(0.0000)	(0.0471)	(0.0000)
Germany	0.4898	0.0357	0.0297	−0.0024	0.0056	0.0015
	(0.2380)	(0.1623)	(0.1233)	(0.0555)	(0.0543)	(0.0367)
Germany (nonzeros only)	0.5531	0.2276	0.3730	−0.0640	0.2055	0.1203
	(0.1700)	(0.3548)	(0.2537)	(0.2900)	(0.2700)	(0.3423)

Table 5.13: Average loadings per country (Bundesland) and for all of Germany for the sparse factor model with $\alpha = 0.05$.

Figures

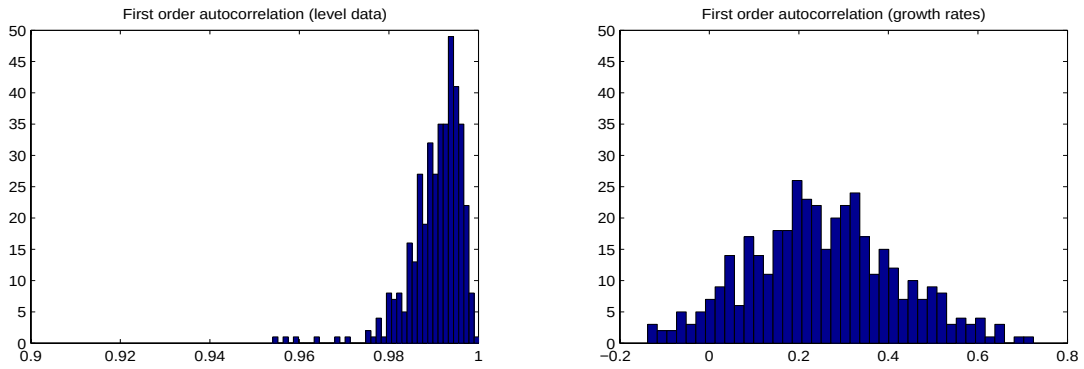

Figure 5.1: Empirical distribution of the first-order autocorrelations for each time series.

Notes: Left panel shows results for the level data, right panel shows results for the growth rates.

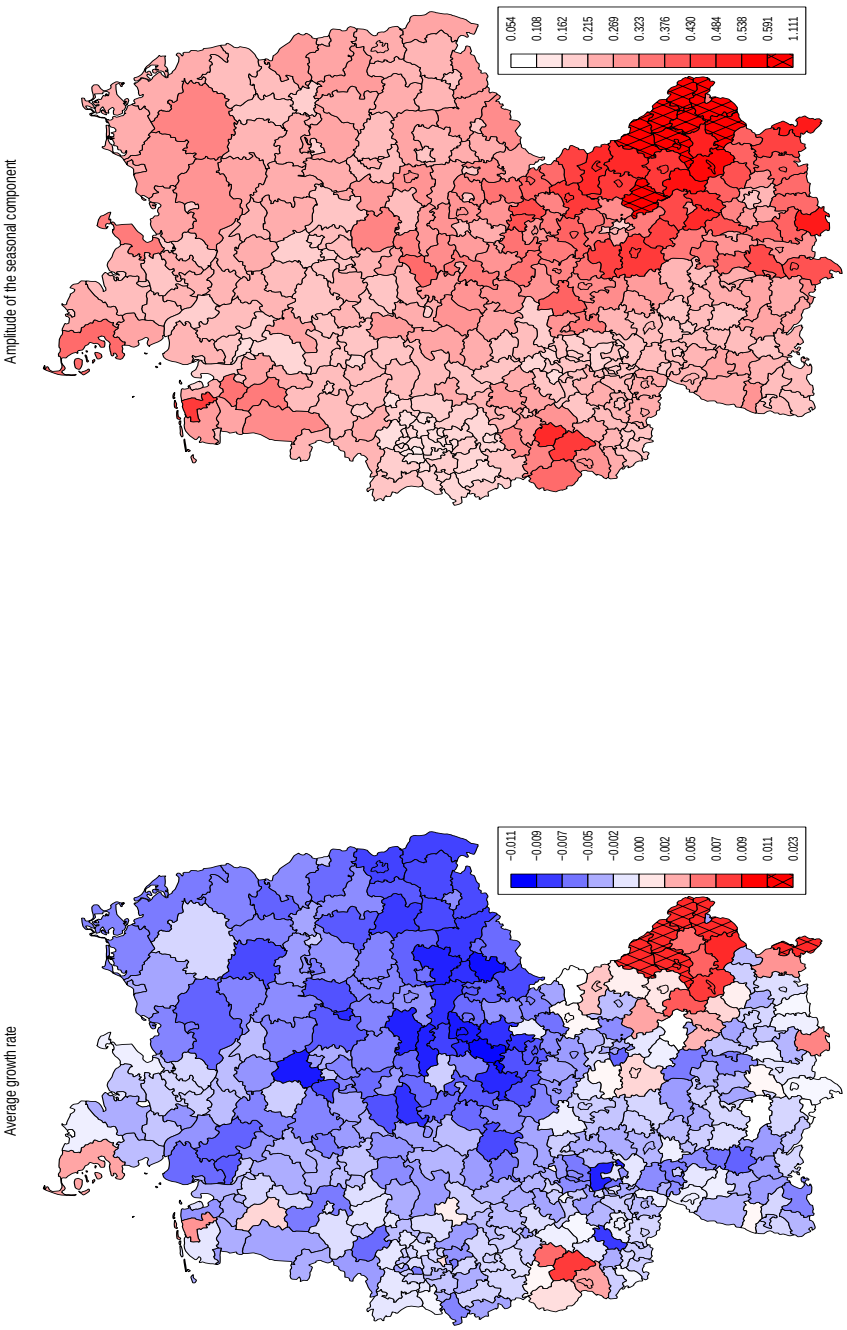

Figure 5.2: Left panel: Average unemployment growth rates of the seasonally adjusted data over the entire period. Right panel: Amplitude of the seasonal pattern extracted from the unemployment growth rates. Hatched areas denote outliers.

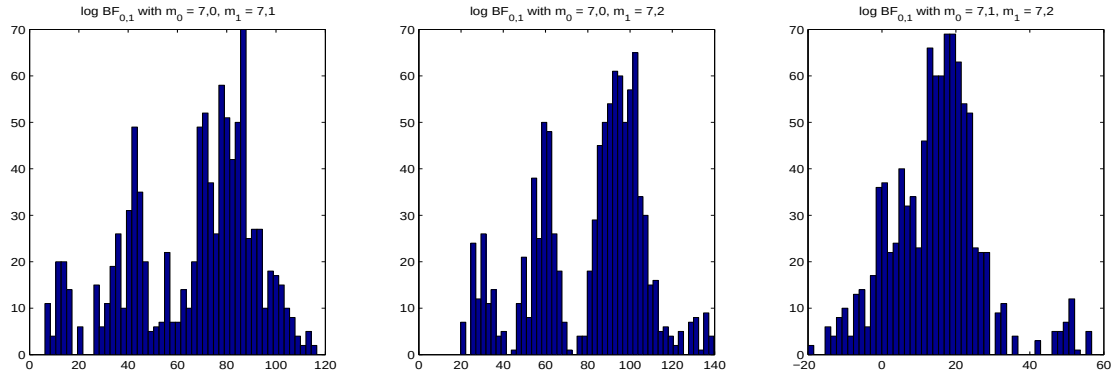

Figure 5.3: Log Bayes Factors comparing the models with $K = 7$, $P = 0$ and $K = 7$, $P = 1$ (left), the models with $K = 7$, $P = 0$ and $K = 7$, $P = 2$ (center), and the models with $K = 7$, $P = 1$ and $K = 7$, $P = 2$ (right).

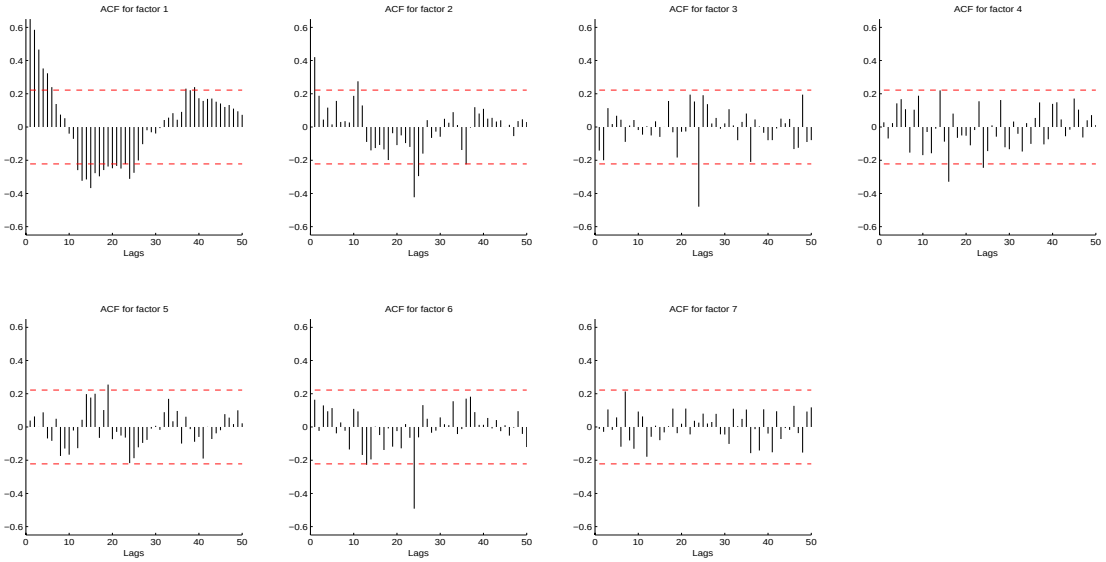

Figure 5.4: ACFs for the factors in the model with $K = 7$, $P = 0$ and $Q = 0$.

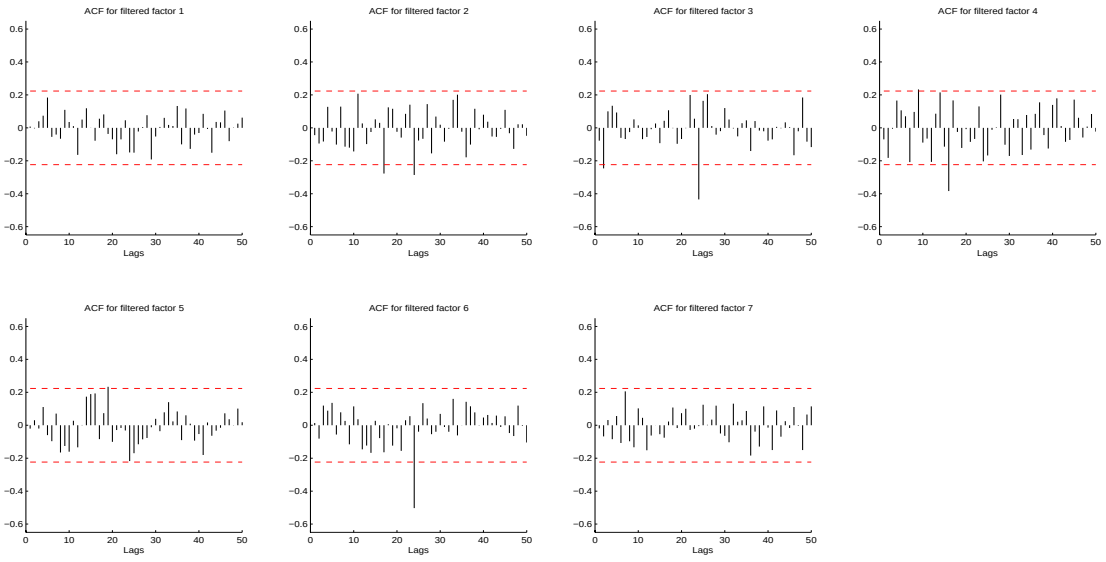

Figure 5.5: ACFs for the filtered factors in the model with $K = 7$, $P = 1$ and $Q = 0$.

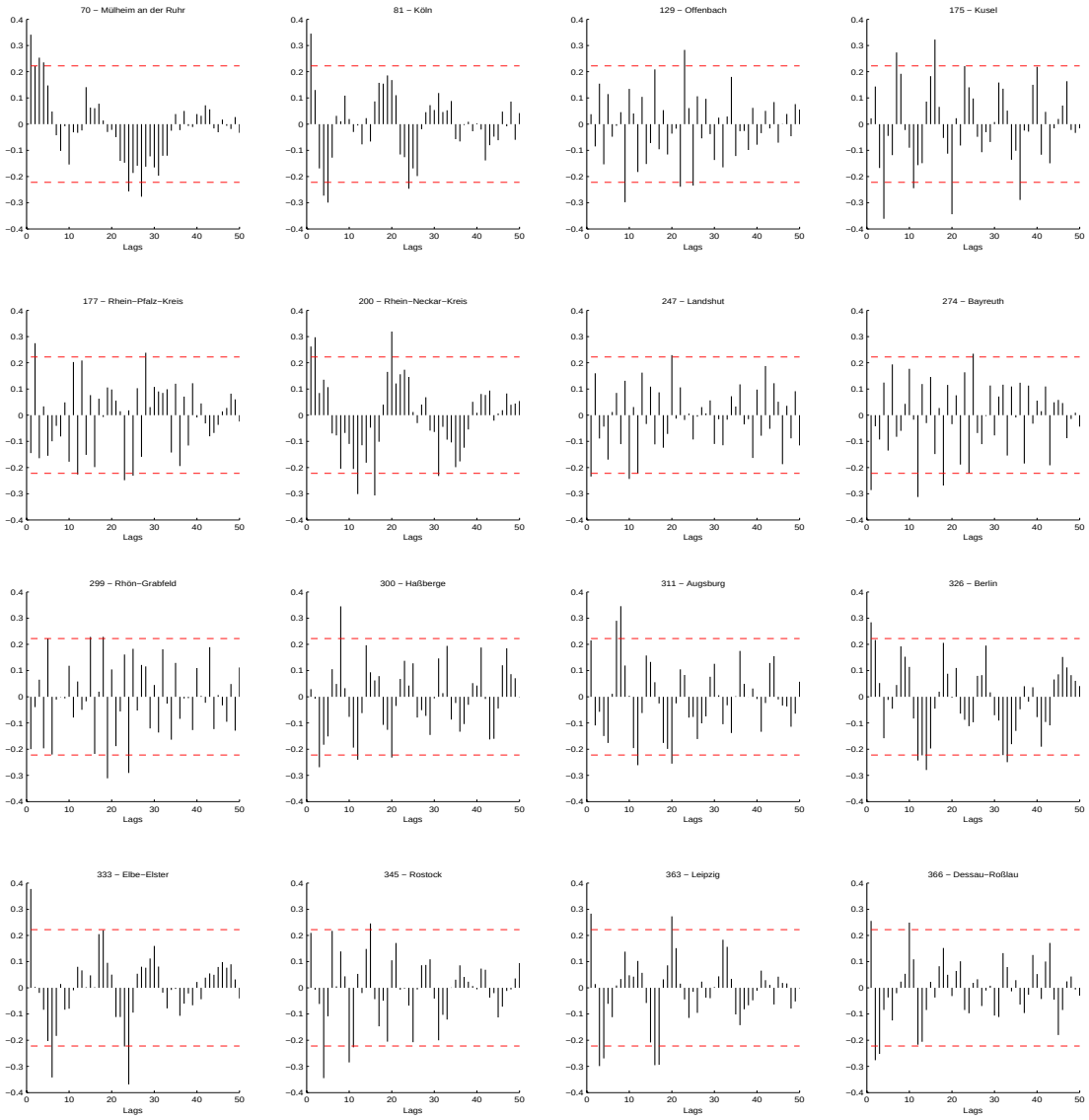

Figure 5.6: Residual ACFs for the model with $K = 7$, $P = 1$ and $Q = 0$ for the 16 counties where more than three out of the 50 estimated autocorrelation coefficients exceed the approximate significance bounds for $\alpha = 0.05$.

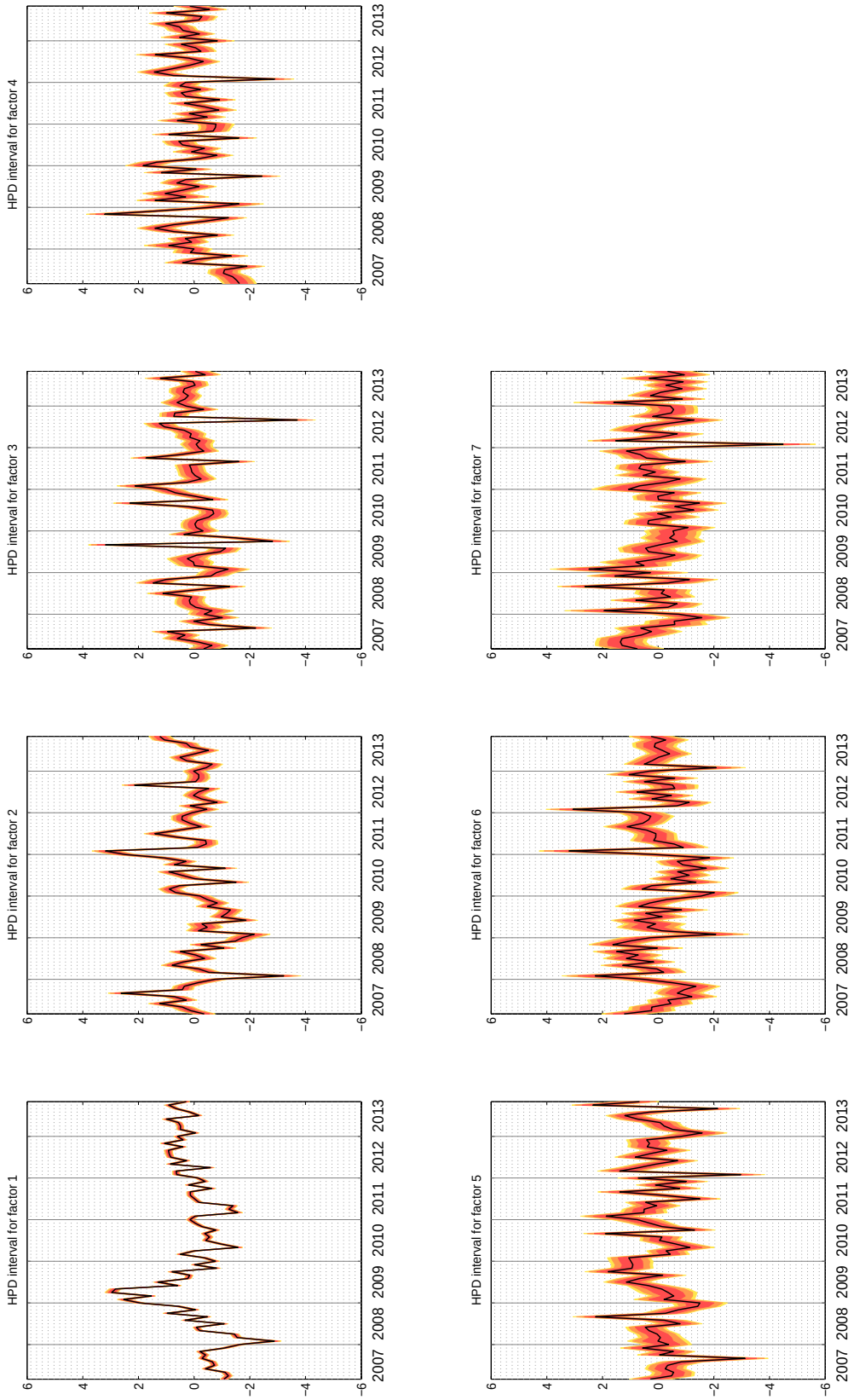

Figure 5.7: Median (black), 68% (red), 90% (orange) and 95% (yellow) highest posterior density intervals of the latent factors for the model with $K = 7$, $P = 1$ and $Q = 0$.

Notes: Factors have been orthogonalized and scaled to unit variance.

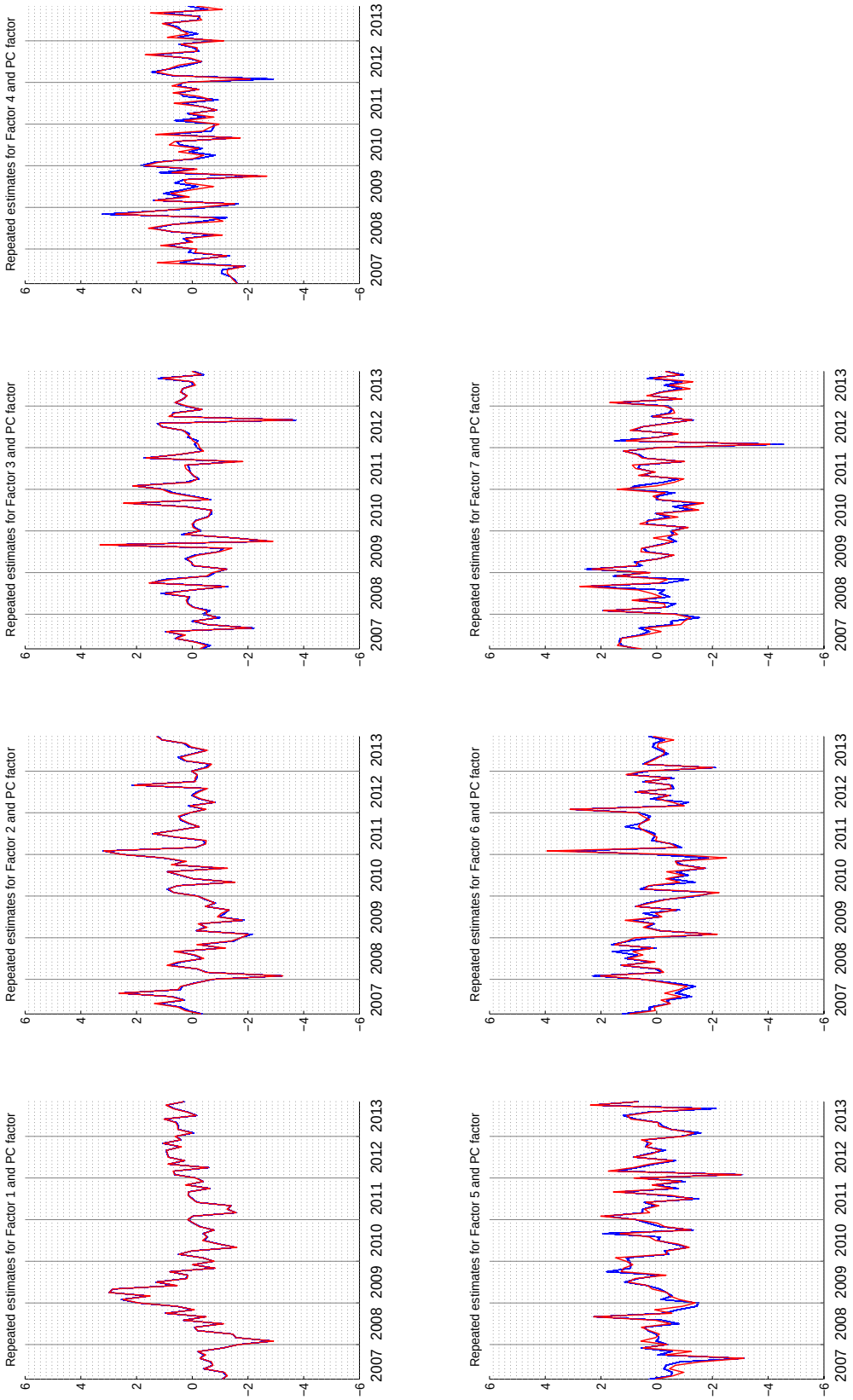

Figure 5.8: 20 repeated estimates of the latent factors for the model with $K = 7$, $P = 1$ and $Q = 0$.

Notes: Factors have been orthogonalized and scaled to unit variance. Red lines show the factors as estimated by PC factor analysis.

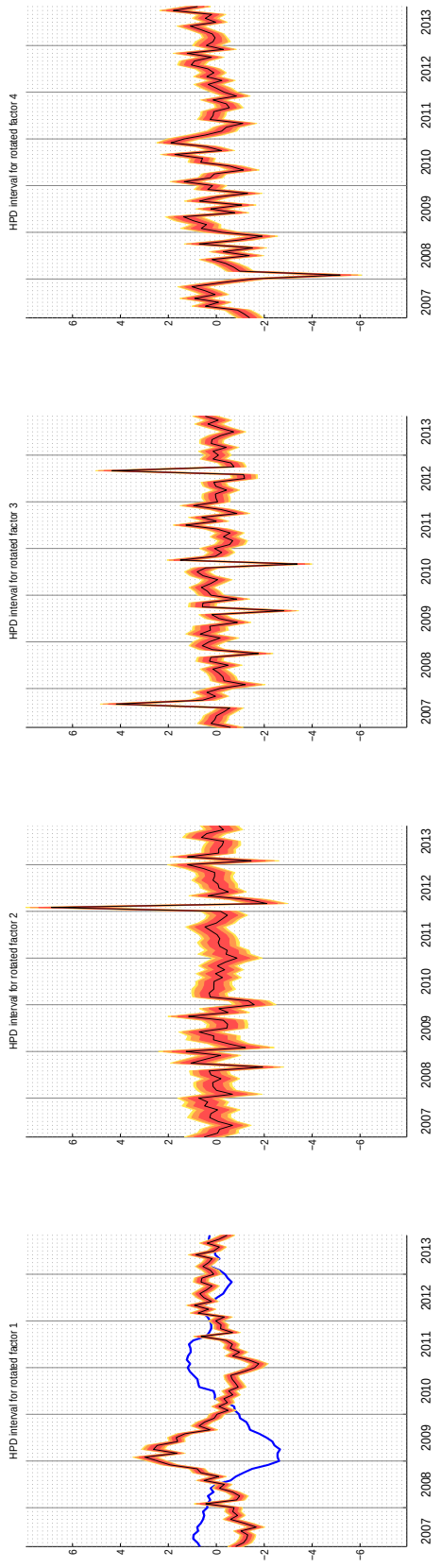

Figure 5.9: Median (black), 68% (red), 90% (orange) and 95% (yellow) highest posterior density intervals of the latent factors for the model with $K = 7$, $P = 1$ and $Q = 0$.

Notes: Factors have been orthogonalized and scaled to unit variance. First factor has been rotated for alignment with the business cycle measured by the ifo index, shown in blue, remaining factors have been rotated to maximize the kurtosis.

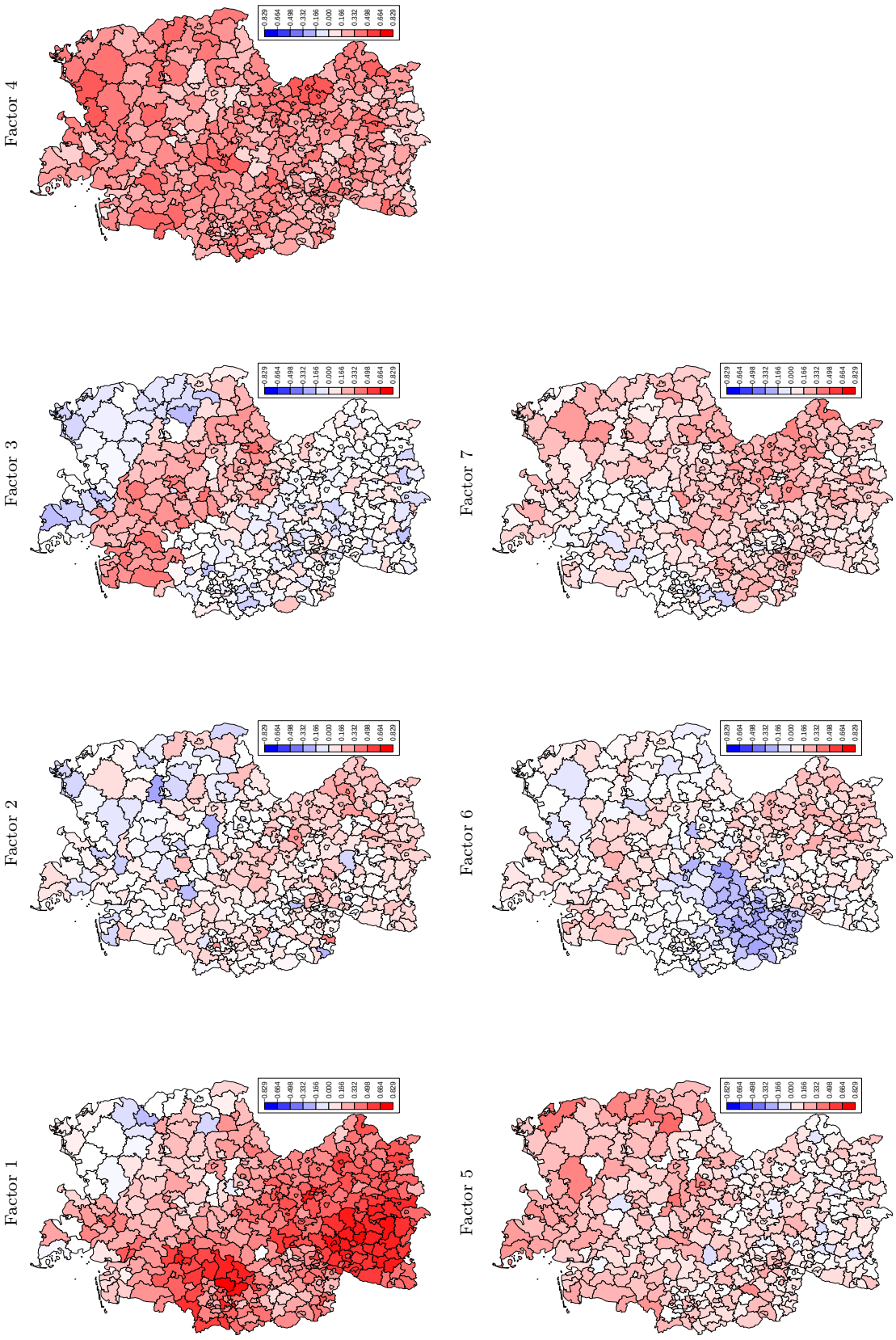

Figure 5.10: Factor loadings for the model with $K = 7$, $P = 1$ and $Q = 0$ with rotated factors.

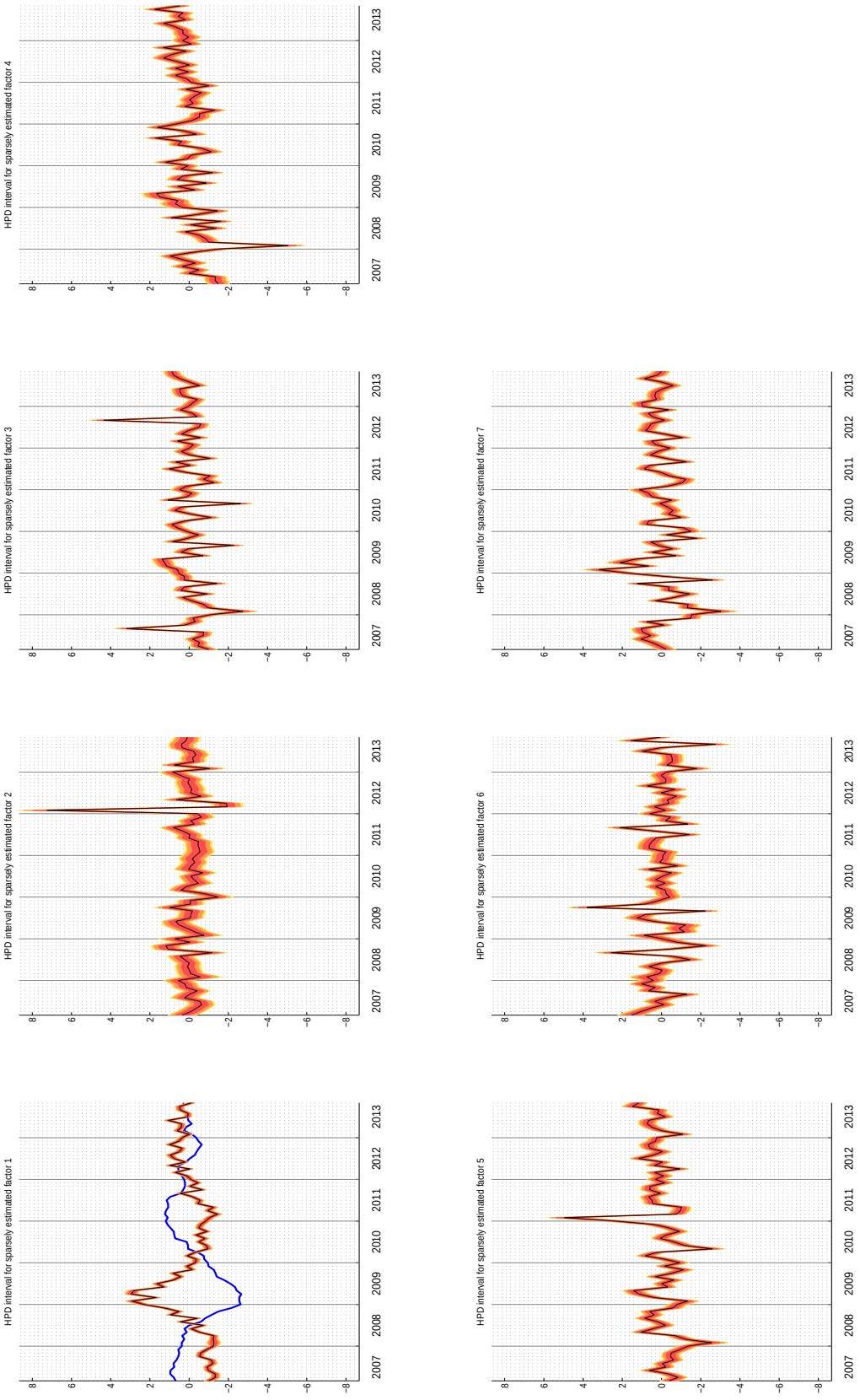

Figure 5.11: Median (black), 68% (red), 90% (orange) and 95% (yellow) highest posterior density intervals of the latent factors for the model with $K = 7$, $P = 1$ and $Q = 0$.

Notes: Sparse factor model using the sparsity structure identified from the rotated factors.

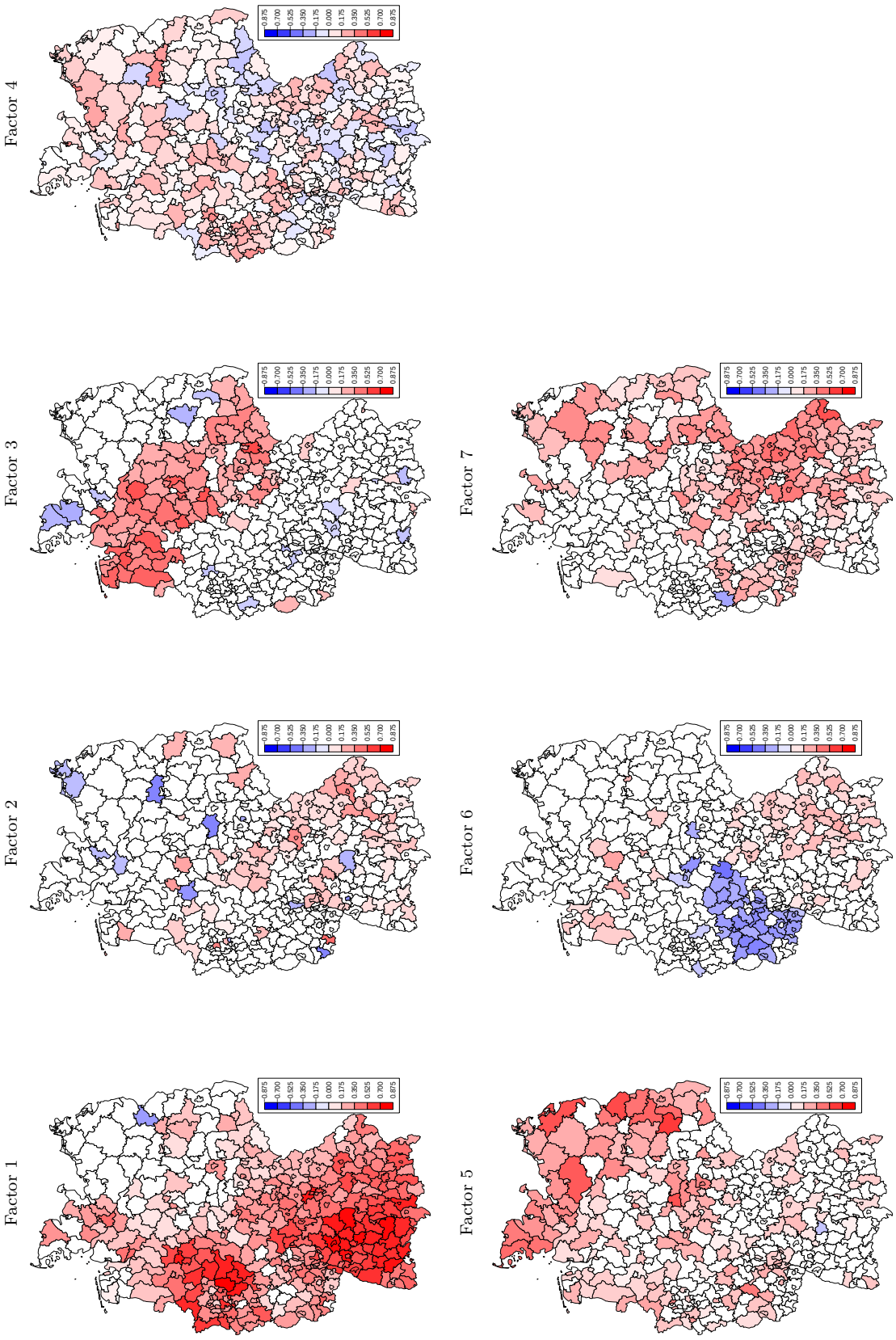

Figure 5.12: Factor loadings for the model with $K = 7$, $P = 1$ and $Q = 0$ with sparse loadings structure based on the rotated factor representation.

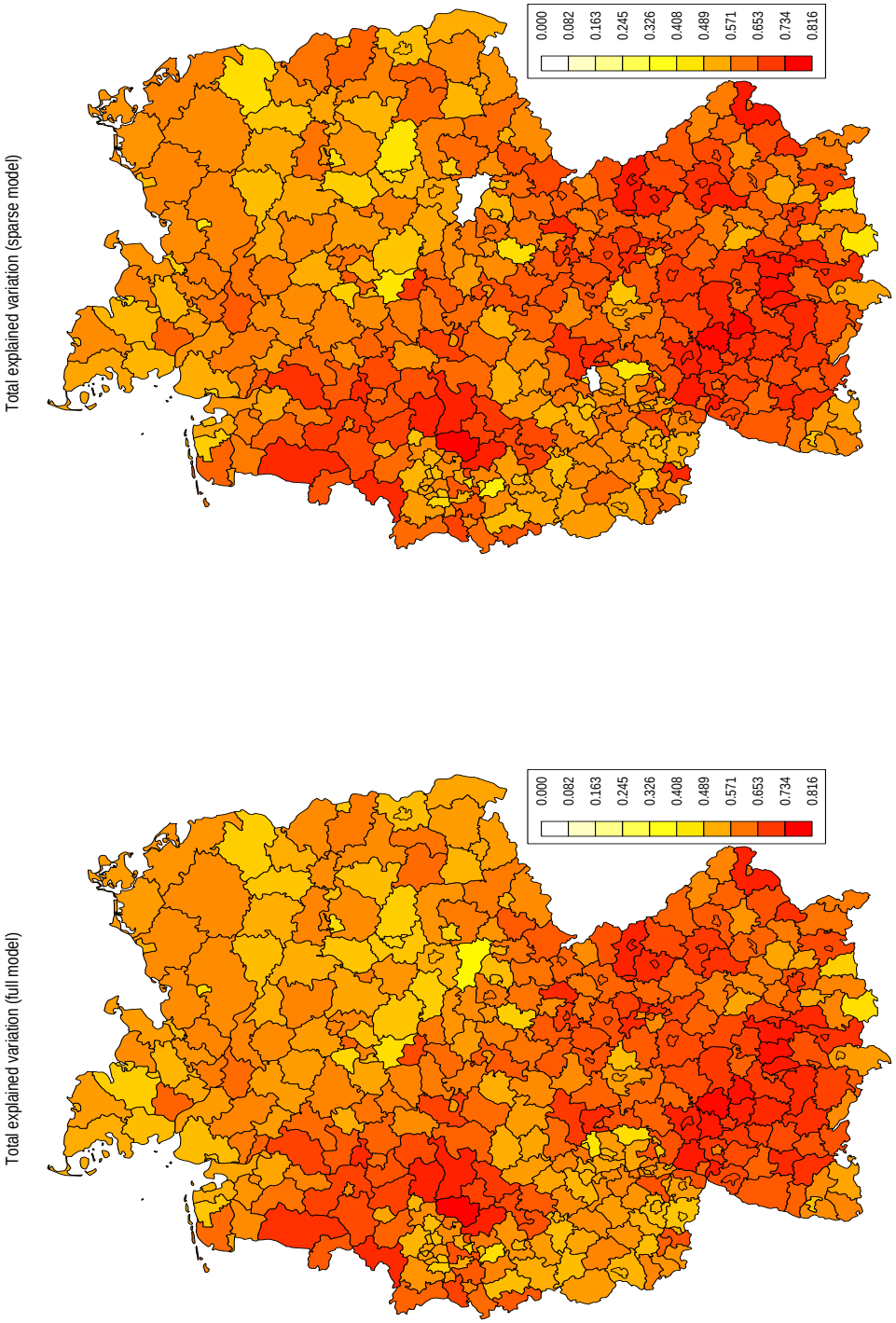

Figure 5.13: Explained variation for the full model (left) and the sparse model (right).

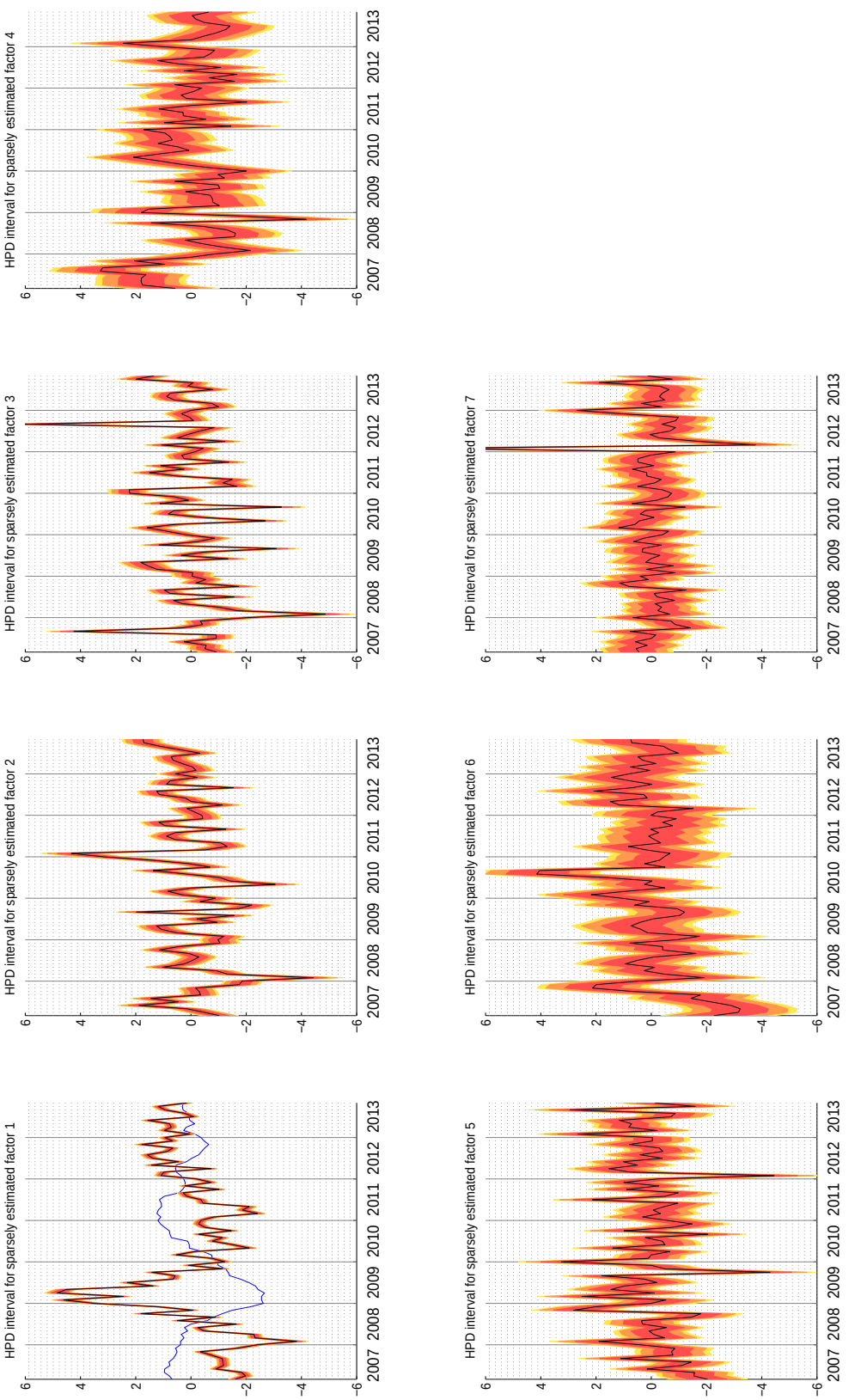

Figure 5.14: Median (black), 68% (red), 90% (orange) and 95% (yellow) highest posterior density intervals of the latent factors for the sparse model with $\alpha = 0.1$.

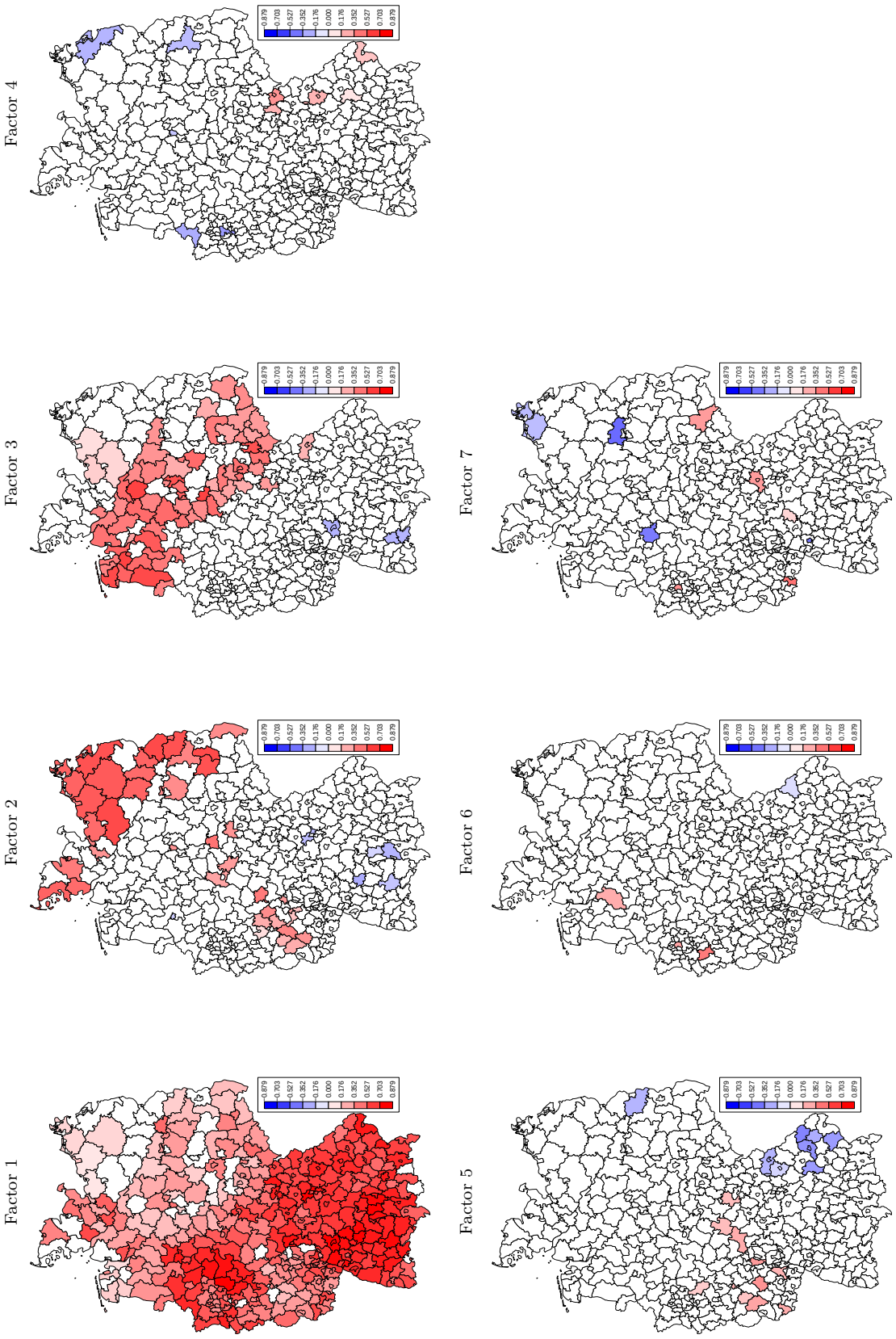

Figure 5.15: Factor loadings for the sparse model with $\alpha = 0.1$.

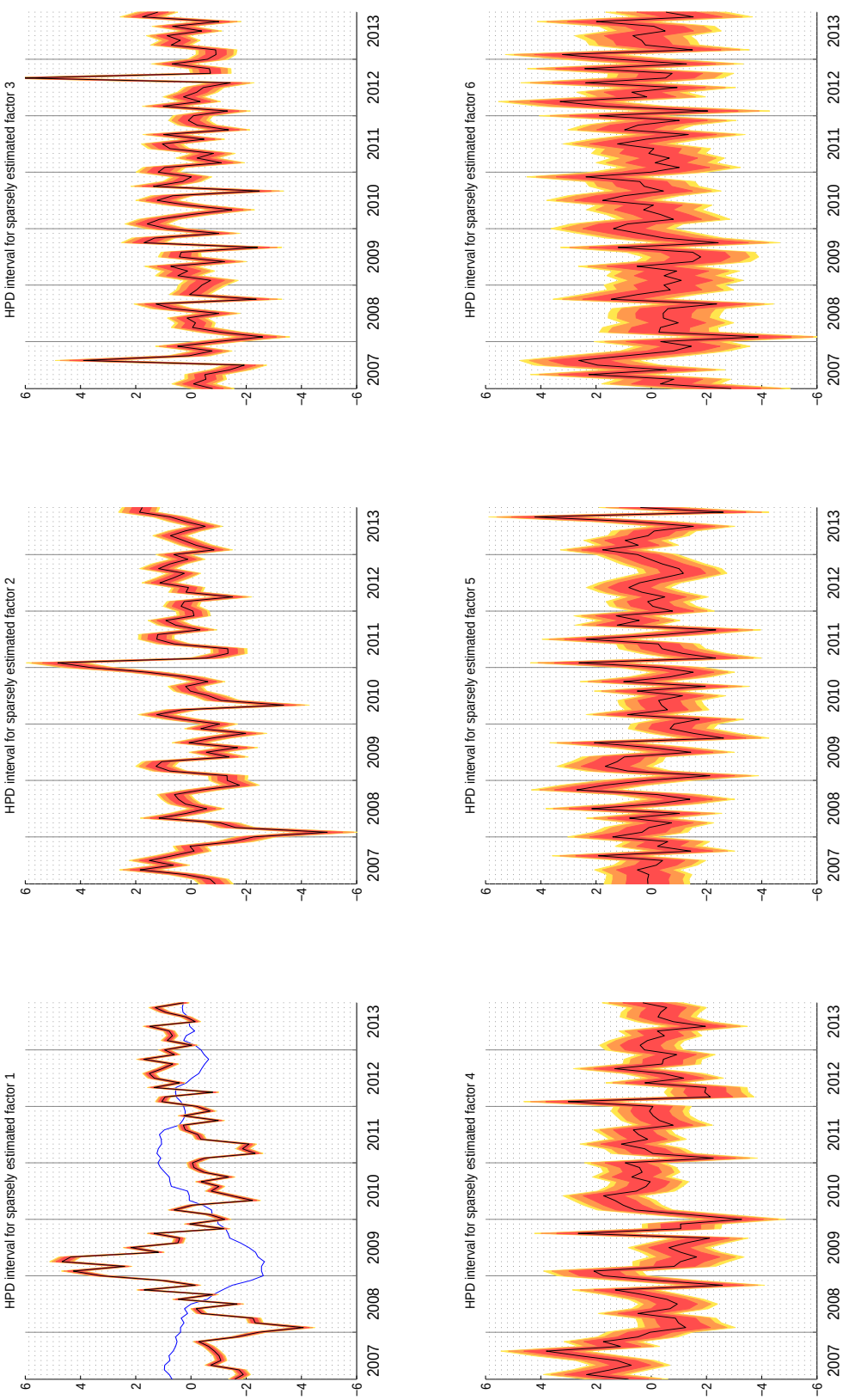

Figure 5.16: Median (black), 68% (red), 90% (orange) and 95% (yellow) highest posterior density intervals of the latent factors for the sparse model with $\alpha = 0.05$.

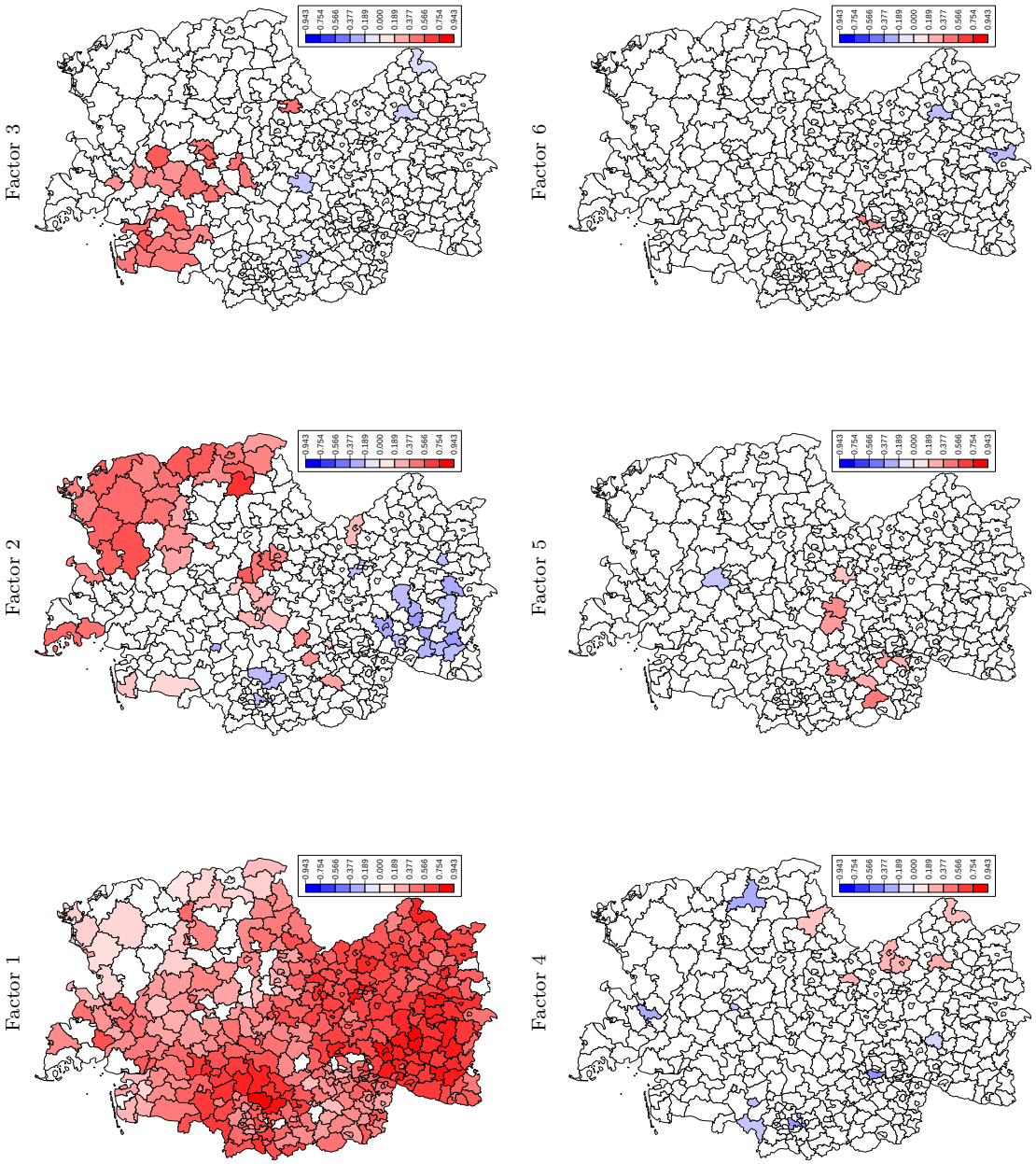

Figure 5.17: Factor loadings for the sparse model with $\alpha = 0.05$.

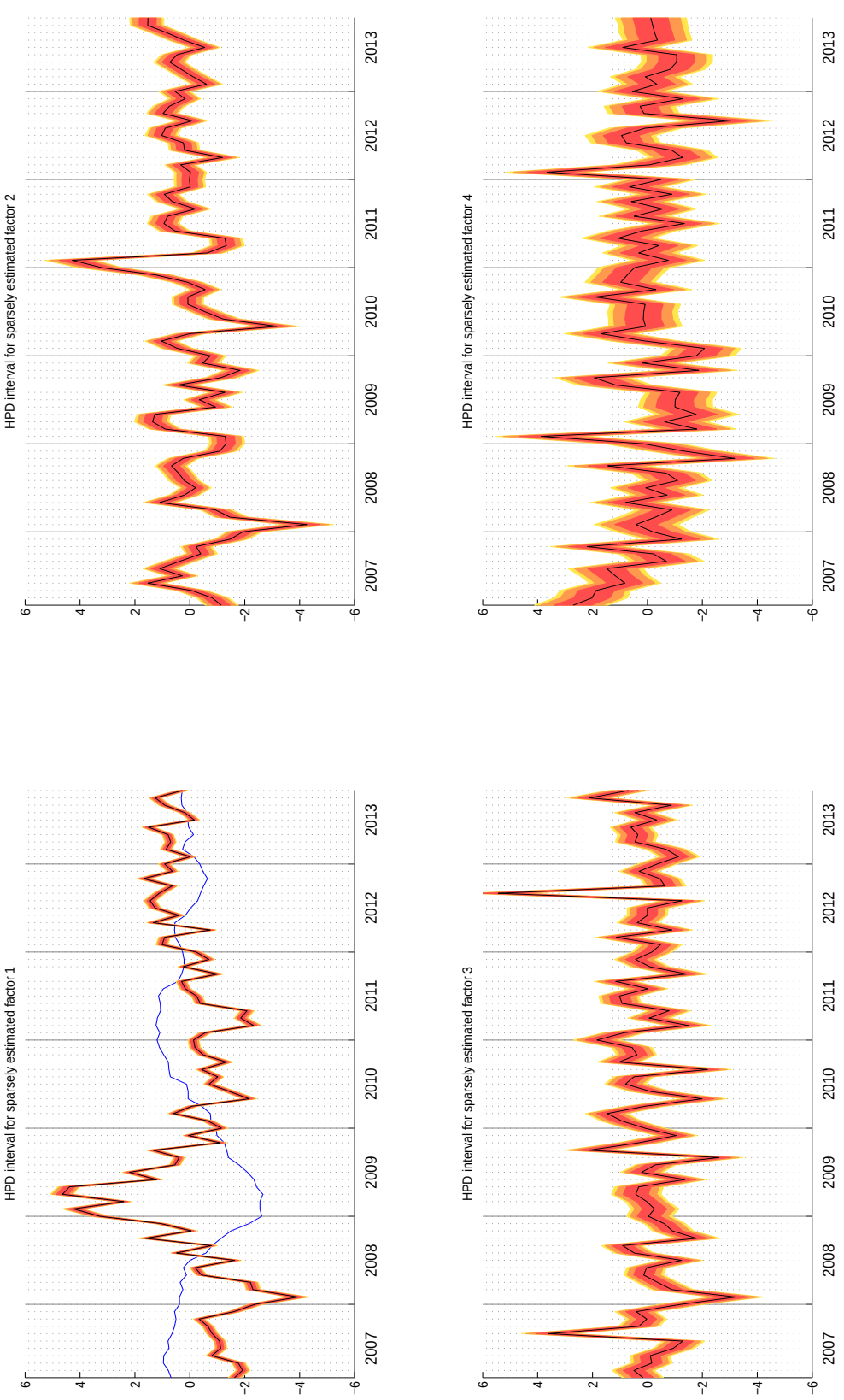

Figure 5.18: Median (black), 68% (red), 90% (orange) and 95% (yellow) highest posterior density intervals of the latent factors for the sparse model with $\alpha = 0.01$.

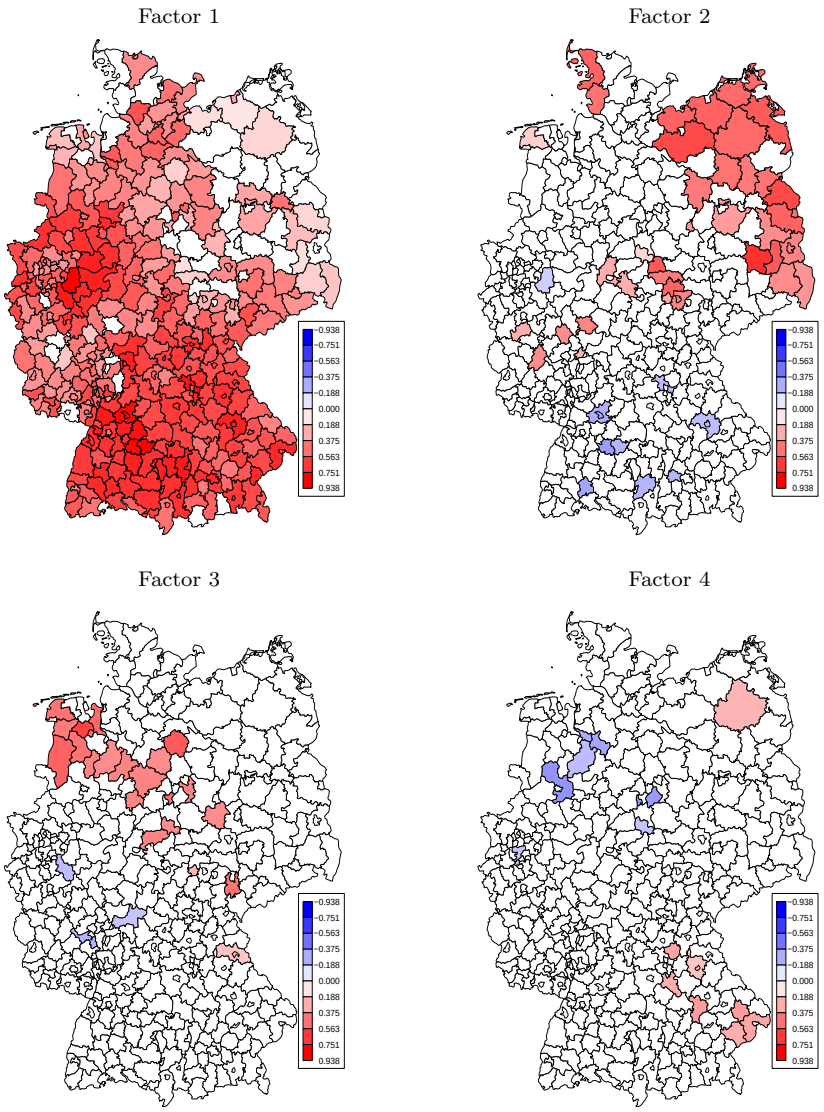

Figure 5.19: Factor loadings for the sparse model with $\alpha = 0.01$.

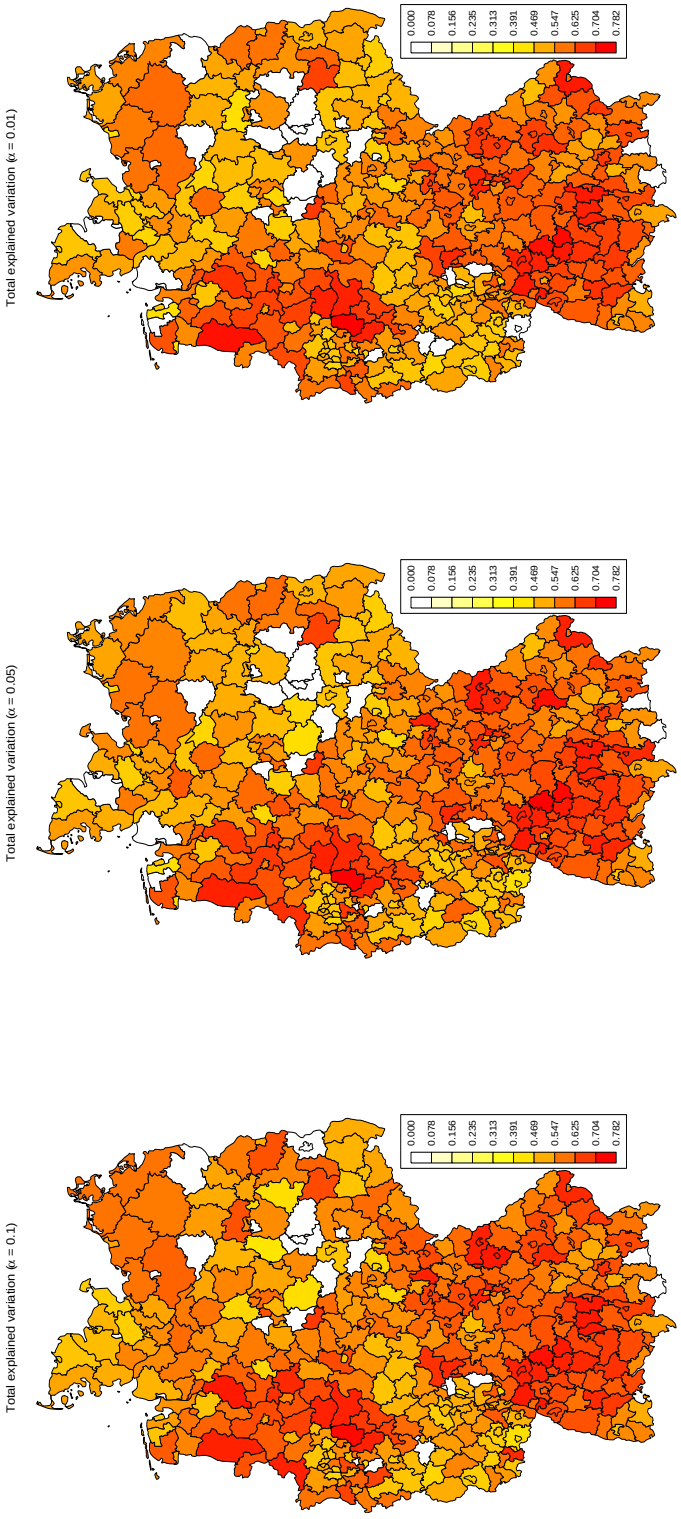

Figure 5.20: Explained variation for the sparse model with $\alpha = 0.1$, $\alpha = 0.05$ and $\alpha = 0.01$.

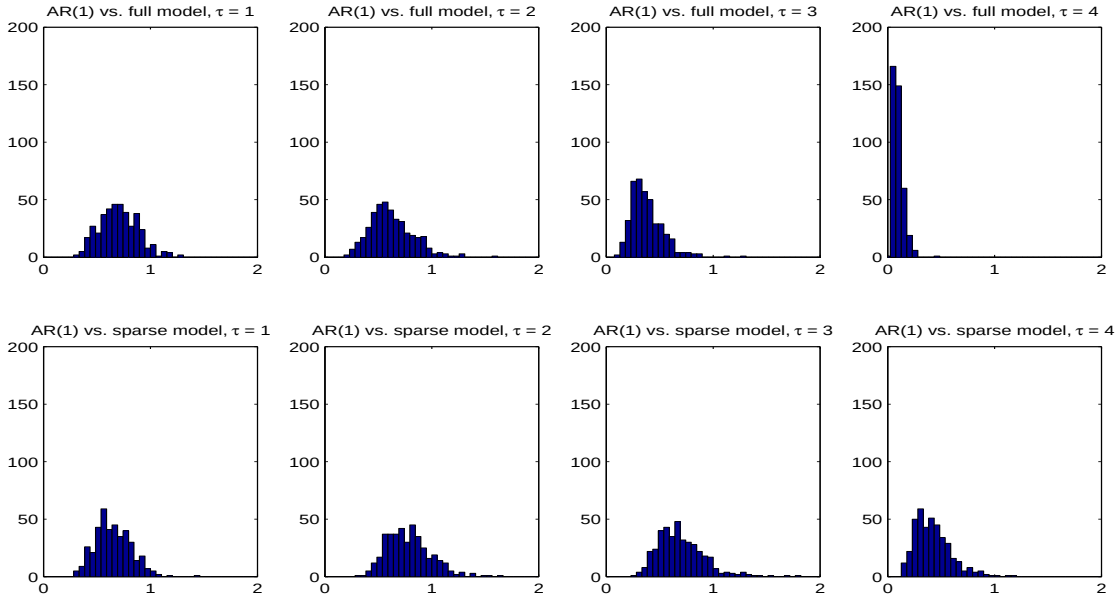

Figure 5.21: Relative RMSFEs for the full (top) and sparse (top) factor models compared to the RMSFEs from a simple AR(1) model.

Appendix 5.A: Full Conditional Distributions for the Unconstrained Gibbs Sampler

The full conditional posterior distributions are then obtained as follows. The loadings are sampled from

$$f(\Lambda|\Sigma,\{\Phi_p\}_{p=1}^P,\{\Theta_q\}_{q=1}^Q,\{f_t\}_{t=1}^T,Y)=\prod_{i=1}^N(2\pi)^{-\frac{K}{2}}|\Omega_{\lambda_i}|^{-\frac{1}{2}} \times\exp\left\{-\frac{1}{2}(\lambda_i-\mu_{\lambda_i})'\Omega_{\lambda_i}^{-1}(\lambda_i-\mu_{\lambda_i})\right\}, \tag{5.70}$$

where

$$\Omega_{\lambda_i}=\left(\frac{1}{\sigma_i^2}\sum_{t=1}^T\theta_i(L)f_t(\theta_i(L)f_t)'+\underline{c}_iI_K\right)^{-1} \tag{5.71}$$

and

$$\mu_{\lambda_i}=\Omega_{\lambda_i}\left(\frac{1}{\sigma_i^2}\sum_{t=1}^T y_{i,t}^*(\theta_i(L)f_t)'\right). \tag{5.72}$$

$\theta_i(L)$ denotes the univariate lag polynomial for the idiosyncratic components of cross-section i, so $y_{i,t}^*=\theta_i(L)y_{i,t}$ and $\theta_i(L)f_t$ denotes the filtered factors, i.e.

$$\theta_i(L)f_t=f_t-\sum_{q=1}^Q\theta_if_{t-q}\quad\text{for } t\in\{Q+1,\dots,T\} \tag{5.73}$$

and

$$[\theta_i(L)[f_1\ \dots\ f_Q]]'=C_i^{-1}[f_1\ \dots\ f_Q]', \tag{5.74}$$

with C_i defined as in Equation (5.20).

The idiosyncratic variances σ_i^2 are sampled independently for each cross-section $i\in\{1,\dots,N\}$ from

$$f(\Sigma|\Lambda,\{\Phi_p\}_{p=1}^P,\{\Theta_q\}_{q=1}^Q,\{f_t\}_{t=1}^T,Y)=\prod_{i=1}^N\frac{b_i^{a_i}}{\Gamma(a_i)}\left(\frac{1}{\sigma_i^2}\right)^{a_i-1}\exp\left\{-\frac{1}{\sigma_i^2}b_i\right\}, \tag{5.75}$$

where $a_i=\frac{1}{2}T+\underline{\alpha}_i$ and $b_i=\frac{1}{2}\sum_{t=1}^T(y_{i,t}^*-\lambda_i'\theta_i(L)f_t)^2+\underline{\beta}_i$.

The P persistence matrices $\{\Phi_p\}_{p=1}^P$ are stacked to $\tilde{\Phi}=[\Phi_1',\ \dots,\ \Phi_P']'$, which is then sampled from

$$f(\tilde{\Phi}|\Sigma,\Lambda,\{\Theta_q\}_{q=1}^Q,\{f_t\}_{t=1}^T,Y)=(2\pi)^{-\frac{KP}{2}}|\Omega_{\tilde{\Phi}}|^{-\frac{1}{2}}$$

$$\times \exp\left\{-\frac{1}{2}(\mathrm{vec}(\tilde{\Phi}) - \mu_{\tilde{\Phi}})' {\Omega_{\tilde{\Phi}}}^{-1} (\mathrm{vec}(\tilde{\Phi}) - \mu_{\tilde{\Phi}})\right\}, \tag{5.76}$$

where $\Omega_{\tilde{\Phi}} = I_K \otimes (\tilde{F}'\tilde{F})^{-1}$ and $\mu_{\tilde{\Phi}} = \mathrm{vec}((\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{F}_{P+1})$, and

$$\tilde{F}_t = [f_t,\ \ldots,\ f_{T-P+(t-1)}]', \tag{5.77}$$

is a shortened $(T-P)\times K$ factor matrix starting at time point t and

$$\tilde{F} = [\tilde{F}_1,\ \ldots,\ \tilde{F}_P], \tag{5.78}$$

see e.g. Ni and Sun (2005).

The persistence parameters in the idiosyncratic error terms are sampled independently for each cross-section $i \in \{1,\ldots,N\}$ from

$$f(\theta_i|\Lambda,\Sigma,\{\Phi_p\}_{p=1}^P,\{f_t\}_{t=1}^T,Y) = \prod_{i=1}^N (2\pi)^{-\frac{K}{2}} |\Omega_{\theta_i}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\theta_i - \mu_{\theta_i})'\Omega_{\theta_i}^{-1}(\theta_i - \mu_{\theta_i})\right\}, \tag{5.79}$$

where

$$\Omega_{\theta_i} = \left(\frac{1}{\sigma_i^2}\tilde{E}_i'\tilde{E}_i + \underline{\Psi}_i\right)^{-1} \tag{5.80}$$

and

$$\mu_{\theta_i} = \Omega_{\theta_i}\left(\frac{1}{\sigma_i^2}\tilde{E}_i'\tilde{e}_{i,Q+1} + \underline{\Psi}_i^{-1}\underline{\zeta}_i\right), \tag{5.81}$$

where

$$\tilde{e}_{i,t} = [e_{i,t},\ \ldots,\ e_{i,T-Q+(t-1)}] \tag{5.82}$$

is a shortened $(T-Q)\times 1$ vector of residuals $e_{i,t} = y_{i,t} - \lambda_i' f_t$, and

$$\tilde{E}_i = [\tilde{e}_{i,1},\ \ldots,\ \tilde{e}_{i,Q}] \tag{5.83}$$

accordingly forms a $(T-Q)\times Q$ matrix.

The latent dynamic factors are obtained by forward-filtering backward-sampling, where the ensemble transform square-root Kalman filter from Tippett et al. (2003) is used. The approach is similar to that in Chapter 3, except that the state and observation equations are Equations (5.25) and (5.23) for the case $S = 0$, i.e.

$$\mathbf{y_t} = \mathbf{\Lambda F_t} + \xi_t, \tag{5.84}$$

and

$$\mathbf{F_t} = \mathbf{H}\mathbf{F_{t-1}} + \epsilon_\mathbf{t}, \tag{5.85}$$

where $\mathbf{\Lambda}$ and $\mathbf{F_t}$ are an $N \times KR$ matrix, and a $KR \times 1$ vector, respectively, and the $\mathbf{\Phi_r}$ in $\mathbf{H}$, as defined in Equation (5.25), become the Φ_r, such that $\mathbf{H}$ is a $KR \times KR$ matrix.

Chapter 6

Conclusion

This thesis discusses issues of model identification in the estimation of static and dynamic factor models by means of Bayesian procedures, focusing on the implications of the rotation problem for Bayesian factor analysis. It implies that the factors and loadings, as well as the parameters governing the factor process in the case of the dynamic factor model, are not uniquely identified. The rotation problem in factor analysis in general has been known and discussed since the beginnings of multiple factor analysis. If the factor model is estimated by principal components, it is not an issue in the estimation process, but only in the interpretation of the results. Thus, rotation techniques are routinely applied to transform the results to obtain better interpretable parameter estimates. In maximum likelihood and Bayesian factor analysis, however, the rotation problem and how it is solved affects the estimation procedure itself. Maximum likelihood factor analysis places constraints on some of the loadings parameters to guarantee unique estimates for the aforementioned model parameters, whereas Bayesian factor analysis uses informative priors to achieve the same. In both cases, however, results may depend on where the constraints or the informative priors are placed, i.e. *how* the rotation problem is solved. This has been found to be an issue in Bayesian factor analysis in the recent years.

In this thesis, I first discuss the implications of underidentified factor models in a general sense, introducing the concept of orthogonal mixing and of orthogonal mixture distributions in Chapter 2. After providing some theoretical foundations, I explain how the issue of label switching known from Markov switching and mixture models and the issue of sign switching, observed e.g. in Bayesian confirmatory factor analysis, are special cases of orthogonal mixing. I introduce an algorithm that is able to remove the mixing from an orthogonally mixed sample. This algorithm uses an orthogonal Procrustes transformation to remove the effect of orthogonal mixing from a sample. It may additionally use weights to account for the different degrees of dispersion in the posterior distributions of the parameters, hence it is called Weighted Orthogonal Procrustes (WOP) algorithm. Next, several samples from different distributions are generated and orthogonal mixing is artificially added to demonstrate how the algorithm works. A second illustration shows how different ways to solve the rotation problem lead to different behavior of the Gibbs sampler in Bayesian static factor analysis.

In Chapter 3, the rotation issue in Bayesian factor analysis is discussed in more depth for static and dynamic factor models. The underidentification of the factor model due to the rotation problem is described in detail. Recent findings in the literature include the ordering problem, i.e. the observation that different orderings of the data produce parameter estimates from different equivalence classes. The different orderings of the data correspond to different sets of constraints, which are albeit all suitable to provide an exact model identification. The constraints, introduced in Bayesian factor analysis in terms of informative prior distributions, are often implemented by enforcing that the loadings matrix is estimated as a positive lower triangular (PLT) matrix. This approach provides a solution to the rotation problem and has routinely been used in the literature, referred to in this thesis as the (ex-ante) PLT approach. As an alternative, a sampler without the aforementioned ex-ante constraints is proposed, whose output is an orthogonally mixed sample. The motivation for this approach can be found in the literature on label switching particularly in the context of Markov switching and mixture models and finds that ex-post identification provides better estimates than imposing constraints ex ante. The WOP algorithm is extended to the case of dynamic factor models and results obtained under the ex-ante PLT identification approach and under the ex-post WOP identification approach are compared for simulated data and for a data set containing 120 macroeconomic time series, for which a dynamic factor model is estimated. The root mean-squared errors of the parameter estimates are generally smaller for the WOP approach than for the PLT approach in the simulation study, and the numerical standard errors are even substantially smaller for the WOP approach than for PLT approach. In the empirical application, it is shown that the WOP approach allows to produce parameter estimates that are invariant to the ordering of the data, hence, the WOP approach is able to overcome the ordering problem.

Chapter 4 discusses the relation between exploratory, confirmatory and sparse factor analysis and some of the pitfalls that may occur in the estimation process. Referring to Chapter 3 and the literature for the first two model types, it is argued that sparse factor analysis may incur multimodality problems. The multimodality issue is briefly demonstrated in a simulation study. Arguing that an ex-post identification approach may also be suitable for sparse factor models, a two-step procedure based on the WOP approach is proposed. The WOP approach does not only provide parameter estimates invariant to the ordering of the data, but also samples from the posterior densities that are likewise invariant to the ordering of the data. As these posterior densities are elliptical for the loadings parameters, it is possible to construct multivariate highest posterior density intervals for the loadings parameters and to find orthogonal transformations that maximize the number of zeros simultaneously contained in these intervals, finding a parsimonious sparse loadings structure. An algorithm is proposed that finds the optimal orthogonal transformation conditional on different criteria, designed e.g. to maximize the number of zero loadings or to find as many zero loadings as possible in a particular column of the loadings matrix. If the number of nonzero loadings in a column becomes too small, the corresponding factor must be assumed to be spurious. The proposed approach is then tried out for simulated data, where the data-generating process is designed

according to previous approaches in the literature. The sparse patterns are overall recognized very well, and the algorithm also turns out to be able both to find out which variables are irrelevant for the factor structure and to find out whether the number of factors in the model has been misspecified, i.e. chosen too large. A comparison to a sampler for sparse Bayesian factor analysis shows that the performance of the proposed two-step approach is similar. In an empirical application, a psychometric data set is analyzed, which has been known to be suitable for the bi-factor model, in which each variable has a high loading on a general factor and an additional nonzero loading on a specific group factor. The bi-factor pattern is recovered quite well by the two-step approach, which tends to find a lower number of factors and an even higher degree of sparsity than implied by the bi-factor model. The results are similar if the sampler for sparse Bayesian factor analysis is used instead.

Chapter 5 analyzes regional labor market data for Germany for 402 counties, which is observed over 82 periods. The data are transformed into growth rates to guarantee stationarity. The dynamic factor model from Chapter 3 is further extended to allow for serial correlation in the error terms. In Bayesian analysis, criteria frequently used for model selection are the deviance information criterion (DIC) and the Bayes factor, which amounts to the marginal likelihood if all models are assumed to be equally likely a priori. It is shown how the marginal likelihood can be obtained with Chib's method for a dynamic factor model. Applying the criteria to the unemployment data set yields a specification with seven factors, one lag in the factor process, and no dynamics in the error terms. These findings are double-checked using additional model diagnostics. It is further shown that the dynamic factors estimated with the WOP approach can be transformed to be highly correlated with the first principal components, which indicates that the WOP approach extracts almost as much information from the data as principal components analysis. The subsequent factor analysis yields a factor that has a high negative correlation with several business-cycle related variables, particularly the ifo business climate indicator. This factor is therefore assumed to be a business climate factor. The remaining factors are harder to interpret. Next, the two-step procedure from Chapter 4 is applied to identify a parsimonious loadings structure and possibly identify spurious factors. Depending on the width of the highest posterior density intervals, the number of non-spurious factors is found to be six or four. The first three factors, which are highly correlated for different parameterizations, show a distinct pattern: Again, the first factor, which has nonzero loadings for most of the counties, is strongly negatively correlated with the ifo business climate indicator. The second factor has nonzero loadings particularly where the first factor has zero loadings, so it describes an alternative cycle, present mainly in the north eastern part of the country. Eventually, the third factor has nonzero loadings predominantly in Lower Saxony, where the first factor likewise has nonzero loadings. Thus it may be interpreted as a complementary cycle. The geographical clustering, in particular the fact that the regional factors closely follow the boundaries of the countries (Bundesländer) may indicate policy-induced cycles. A forecasting exercise shows that both factor models are outperformed by AR(1) models estimated separately for each series, however, the sparse

factor model performs substantially better for longer forecast horizons than the full factor model.

Altogether, this thesis proposes the use of the suggested WOP ex-post identification approach for the estimation of static and dynamic factor models with full or sparse loadings matrices, demonstrates the advantages of the approach in a number of simulation studies and performs several empirical analyses that underline the findings.

Bibliography

Aguilar, O. and West, M. (2000). Bayesian Dynamic Factor Models and Portfolio Allocation. *Journal of Business & Economic Statistics*, 18(3):338–357.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Amengual, D. and Watson, M. W. (2007). Consistent Estimation of the Number of Dynamic Factors in a Large N and T Panel. *Journal of Business & Economic Statistics*, 25:91–96.

Anderson, T. (1963). The Use of Factor Analysis in the Statistical Analysis of Multiple Time Series. *Psychometrika*, 28(1):1–25.

Anderson, T., Olkin, I., and Underhill, L. (1987). Generation of Random Orthogonal Matrices. *SIAM Journal on Scientific and Statistical Computing*, 8:625–629.

Anderson, T. and Rubin, H. (1956). *Statistical Inference in Factor Models*, volume 5 of *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 111–150. University of California Press.

Artin, M. (1991). *Algebra*. Prentice Hall, Englewood Cliffs.

Artis, M., Dreger, C., and Kholodilin, K. (2011). What Drives Regional Business Cycles? The Role of Common and Spatial Components. *The Manchester School*, 79(5):1035–1044.

Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2012). The Directional Identification Problem in Bayesian Factor Analysis: An Ex-Post Approach. Kiel Working Papers 1799, Kiel Institute for the World Economy.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70:191–221.

Bai, J. and Ng, S. (2007). Determining the Number of Primitive Shocks in Factor Models. *Journal of Business and Economics Statistics*, 25(1):52–60.

Bai, J. and Ng, S. (2013). Principal Components Estimation and Identification of Static Factors. *Journal of Econometrics*, 176(1):18–29.

Bai, J. and Wang, P. (2012). Identification and Estimation of Dynamic Factor Models. MPRA Paper 38434, University Library of Munich, Germany.

Banbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian Vector Autoregressions. *Journal of Applied Econometrics*, 25(1):71–92.

Barhoumi, K., Darné, O., and Ferrara, L. (2013). Dynamic Factor Models: A Review of the Literature . Working papers 430, Banque de France.

Bartholomew, D. (1984). The Foundations of Factor Analysis. *Biometrika*, 71(2):221–232.

Bekker, P. A. (1986). A Note on Identification of Restricted Factor Loading Matrices. *Psychometrika*, 51(4):607–611.

Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Bentler, P. and Tanaka, J. (1983). Problems with EM algorithms for ML Factor Analysis. *Psychometrika*, 48(2):247–251.

Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107(1):238–246.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

Bernanke, B., Boivin, J., and Eliasz, P. S. (2005). Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach. *The Quarterly Journal of Economics*, 120(1):387–422.

Bernstein, D. (2009). *Matrix Mathematics*. Princeton University Press, Princeton, NJ.

Besag, J. (1989). A Candidate's Formula: A Curious Result in Bayesian Prediction. *Biometrika*, 76(1):183.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian Infinite Factor Models. *Biometrika*, 98(2):291–306.

Blanchard, O. J. and Katz, L. F. (1992). Regional Evolutions. *Brookings Papers on Economic Activity*, 23(1):1–76.

Boivin, J. and Giannoni, M. (2006). DSGE Models in a Data-Rich Environment. NBER Technical Working Papers 0332, National Bureau of Economic Research, Inc.

Boivin, J. and Ng, S. (2006). Are More Data Always Better for Factor Analysis? *Journal of Econometrics*, 132(1):169–194.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, New York.

Box, G. and Draper, N. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, Oxford.

Boysen-Hogrefe, J. and Pape, M. (2011). More than Just One Labor Market Cycle in Germany? : An Analysis of Regional Unemployment Data. *Journal for Labour Market Research - Zeitschrift für ArbeitsmarktForschung*, 44(3):279–292.

Bozdogan, H. and Ramirez, D. E. (1987). An Expert Model Selection Approach to Determine the "Best" Pattern Structure in Factor Analysis Models. In Bozdogan, H. and Gupta,

A., editors, *Multivariate Statistical Modelling and Data Analysis*, pages 35–60. D. Reidel Publishing.

Bozdogan, H. and Shigemasu, K. (1998). Bayesian Factor Analysis Model and Choosing the Number of Factors Using a New Informational Complexity Criterion. In Rizzi, A., Vichi, M., and Bock, H.-H., editors, *Advances in Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 335–342. Springer Berlin Heidelberg.

Breitung, J. and Pigorsch, U. (2013). A Canonical Correlation Approach for Selecting the Number of Dynamic Factors. *Oxford Bulletin of Economics and Statistics*, 75(1):23–36.

Brillinger, D. (1964). A Frequency Approach to the Techniques of Principal Components, Factor Analysis and Canonical Covariates in the Case of Stationary Time Series. Invited paper, Royal Statistical Society Conference.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco.

Burns, A. F. and Mitchell, W. C. (1946). *Measuring Business Cycles*. NBER Books. National Bureau of Economic Research, Inc.

Carneiro, P., Hansen, K. T., and 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 Choice. *International Economic Review*, 44(2):361–422.

Carter, C. K. and Kohn, R. (1994). On Gibbs Sampling for State Space Models. *Biometrika*, 81(3):541–553.

Carvalho, C. M. (2006). Structure and sparsity in high-dimensional multivariate analysis. PhD Thesis, Department of Statistical Science, Duke University.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(4):1438–1456.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Cattell, R. (1952). *Factor Analysis*. Harper, New York.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276.

Celeux, G. (1998). Bayesian Inference for Mixtures: The Label-Switching Problem. In Payne, R. and Green, P. J., editors, *COMPSTAT 98—Proc. in Computational Statistics*, pages 227–233.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica*, 51(5):1281–1304.

Chan, J., Leon-Gonzalez, R., and Strachan, R. W. (2013). Invariant Inference and Efficient Computation in the Static Factor Model. CAMA Working Paper 32/2013, Centre for Applied Macroeconomic Analysis.

Chan, J. C. C. and Jeliazkov, I. (2009). Efficient Simulation and Integrated Likelihood Estimation in State Space Models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1:101–120.

Charles, D. (1998). Constrained PCA Techniques for the Identification of Common Factors in Data. *Neurocomputing*, 22(1):145–156.

Cheng, M.-Y., Hall, P., and Turlach, B. A. (1999). High-derivative parametric enhancements of nonparametric curve estimators. *Biometrika*, 86(2):417–428.

Cheung, G. W. and Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2):233–255.

Chib, S. (1995). Marginal Likelihood From The Gibbs Output. *Journal of the American Statistical Association*, 90:1313–1321.

Chib, S. and Greenberg, E. (1994). Bayes Inference in Regression Models with ARMA (p, q) Errors. *Journal of Econometrics*, 64(1–2):183–206.

Clarke, B., Andrew, and Barron, R. (1990). Information-theoretic Asymptotics of Bayes Methods. *IEEE Transactions on Information Theory*, 36:453–471.

Connor, G. and Korajczyk, R. A. (1986). Performance Measurement With the Arbitrage Pricing Theory : A New Framework for Analysis. *Journal of Financial Economics*, 15(3):373–394.

Connor, G. and Korajczyk, R. A. (1988). Risk and Return in an Equilibrium APT : Application of a New Test Methodology. *Journal of Financial Economics*, 21(2):255–289.

Connor, G. and Korajczyk, R. A. (1993). A Test for the Number of Factors in an Approximate Factor Model. *Journal of Finance*, 48(4):1263–1291.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., and Piatek, R. (2014). Bayesian Exploratory Factor Analysis. *Journal of Econometrics*, 183(1):31–57.

Dawid, A. P. (1981). Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application. *Biometrika*, 68(1):265–274.

De Grauwe, P. (2010). The Scientific Foundation of Dynamic Stochastic General Equilibrium (DSGE) Models. *Public Choice*, 144(3-4):413–443.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, SERIES B*, 39(1):1–38.

Diebold, F. X. and Rudebusch, G. D. (1996). Measuring Business Cycles: A Modern Perspective. *The Review of Economics and Statistics*, 78(1):67–77.

Doz, C., Giannone, D., and Reichlin, L. (2011). A Two-Step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering. *Journal of Econometrics*, 164(1):188–205.

Doz, C., Giannone, D., and Reichlin, L. (2012). A Quasi-Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics*, 94(4):1014–1024.

Dunn, J. (1973). A Note on a Sufficiency Condition for Uniqueness of a Restricted Factor Matrix. *Psychometrika*, 38(1):141–143.

Dutilleul, P. (1999). The MLE Algorithm for the Matrix Normal Distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123.

Eickmeier, S. (2005). Common Stationary and Non-Stationary Factors in the Euro Area Analyzed in a Large-Scale Factor Model. Discussion Paper Series 1: Economic Studies 2005,02, Deutsche Bundesbank, Research Centre.

Eickmeier, S. and Breitung, J. (2006). How Synchronized are New EU Member States With the Euro Area? Evidence From a Structural Factor Model. *Journal of Comparative Economics*, 34(3):538–563.

Engle, R. and Watson, M. (1981). A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates. *Journal of the American Statistical Association*, 76(376):774–781.

Erosheva, E. A. and Curtis, S. M. (2013). Dealing with Rotational Invariance in Bayesian Confirmatory Factor Analysis. Technical Report 589, Department of Statistics, University of Washington.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., and Chen, B.-C. (1998). New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program. *Journal of Business and Economic Statistics*, 16(2):127–152.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The Generalized Dynamic-Factor Model: Identification And Estimation. *The Review of Economics and Statistics*, 82(4):540–554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2004). The Generalized Dynamic Factor Model: Consistency and Rates. *Journal of Econometrics*, 119(2):231–255.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association*, 100:830–840.

Forni, M. and Lippi, M. (1997). *Aggregation and the Microfoundations of Dynamic Macroeconomics*. Oxford University Press, Oxford.

Forni, M. and Lippi, M. (2001). The Generalized Dynamic Factor Model: Representation Theory. *Econometric Theory*, 17(06):1113–1141.

Francis, N., Owyang, M. T., and Özge Savascin (2012). An Endogenously Clustered Factor Approach to International Business Cycles. Working Papers 2012-014, Federal Reserve Bank of St. Louis.

Frisch, R. (1934). *Statistical Confluence Analysis by Means of Complete Regression Systems*. Universitetets Økonomiske Instituut Publikasjon. Universitetets Økonomiske Instituut.

Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association*, 96(453):194–209.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.

Frühwirth-Schnatter, S. and Lopes, H. F. (2012). Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown. Technical report, University of Chicago Booth School of Business.

Gelfand, A. and Smith, A. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85:398–409.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Geweke, J. (1977). The Dynamic Factor Analysis of Economic Time Series. In Aigner, D. and Goldberger, A., editors, *Latent Variables in Socio-Economic Models*, pages 365–382. North Holland, Amsterdam.

Geweke, J. (1991). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. Staff Report 148, Federal Reserve Bank of Minneapolis.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments. In *Bayesian Statistics 4*, pages 169–193. Oxford University Press.

Geweke, J. and Zhou, G. (1996). Measuring the Pricing Error of the Arbitrage Pricing Theory. *Review of Financial Studies*, 9(2):557–587.

Geweke, J. F. and Singleton, K. J. (1981). Maximum Likelihood "Confirmatory" Factor Analysis of Economic Time Series. *International Economic Review*, 22(1):37–54.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, Boca Raton, USA.

Glanz, H. and Carvalho, L. (2013). An Expectation-Maximization Algorithm for the Matrix Normal Distribution. *ArXiv e-prints*.

Goldberger, A. S. (1972). Structural Equation Methods in the Social Sciences. *Econometrica*, 40(6):979–1001.

Golub, G. H. and van Loan, C. F. (2013). *Matrix Computations*. The Johns Hopkins University Press, 4th edition edition.

Gorsuch, R. L. (1983). *Factor Analysis*. Lawrence Erlbaum Associates, New Jersey.

Green, B. F. (1952). The Orthogonal Approximation of an Oblique Simple Structure in Factor Analysis. *Psychometrika*, 17:429–440.

Grün, B. and Leisch, F. (2009). Dealing With Label Switching in Mixture Models Under Genuine Multimodality. *Journal of Multivariate Analysis*, 100(5):851–861.

Hallin, M. and Liska, R. (2007). Determining the Number of Factors in the General Dynamic Factor Model. *Journal of the American Statistical Association*, 102:603–617.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton NJ.

Hamilton, J. D. and Owyang, M. T. (2012). The Propagation of Regional Recessions. *The Review of Economics and Statistics*, 94(4):935–947.

Hannan, E. J. and Deistler, M. (2012). *The Statistical Theory of Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Hanson, T. and McMillan, G. T. (2012). Scheffé Style Simultaneous Credible Bands for Regression Surfaces with Application to Ache Honey Gathering. *Journal of Data Science*, 10(1):175–193.

Hauck, W. W. (1983). A Note on Confidence Bands for the Logistic Response Curve. *The American Statistician*, 37(2):158–160.

Heckman, J. J., Humphries, J. E., Veramendi, G., and Urzua, S. S. (2014). Education, Health and Wages. NBER Working Papers 19971, National Bureau of Economic Research, Inc.

Heckman, J. J. and Pinto, R. (2013). Causal Analysis after Haavelmo. NBER Working Papers 19453, National Bureau of Economic Research, Inc.

Hendry, D. F. and Doornik, J. A. (1994). Modelling Linear Dynamic Econometric Systems. *Scottish Journal of Political Economy*, 41(1):1–33.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer, New York, NY, 1st edition.

Holzinger, K. J. and Swineford, F. (1937). The Bi-factor Method. *Psychometrika*, 2(1):41–54.

Holzinger, K. J. and Swineford, F. (1939). A Study in Factor A: The Stability of a Bi-Factor Solution. Technical Report 48, The University of Chicago Press.

Howe, W. G. (1955). Some Contributions to Factor Analysis. Technical Report ORNL-1919, Oak Ridge National Laboratory.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, New York.

Ishwaran, H. and Rao, J. S. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, 33(2):730–773.

Izenman, A. J. (1975). Reduced-Rank Regression for the Multivariate Linear Model. *Journal of Multivariate Analysis*, 5(2):248–264.

Jacobs, J. P. A. M. and Otter, P. W. (2008). Determining the Number of Factors and Lag Order in Dynamic Factor Models: A Minimum Entropy Approach. *Econometric Reviews*, 27(4–6):385–397.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1):50–67.

Jazwinski, A. (1970). *Stochastic Processes and Filtering Theory*. Academic Press, New York.

Jennrich, R. I. (1978). Rotational Equivalence of Factor Loading Matrices with Prespecified Values. *Psychometrika*, 43:421–426.

Johannsen, W. (1911). The Genotype Conception of Heredity. *The American Naturalist*, 45(531):129–159.

Jöreskog, K. (1967). Some Contributions to Maximum Likelihood Factor Analysis. *Psychometrika*, 32(4):443–482.

Jöreskog, K. (1969). A General Approach to Confirmatory Maximum Likelihood Factor Analysis. *Psychometrika*, 34(2):183–202.

Jöreskog, K. (1971). Statistical Analysis of Sets of Congeneric Tests. *Psychometrika*, 36(2):109–133.

Jöreskog, K. G. (1979a). Author's Addendum to: A General Approach to Confirmatory Factor Analysis. In Magidson, J., editor, *Advances in Factor Analysis and Structural Equation Models*. Abt Books, Cambridge, MA.

Jöreskog, K. G. (1979b). Structural Equations Models in the Social Sciences: Specification, Estimation and Testing. In Magidson, J., editor, *Advances in Factor Analysis and Structural Equation Models*. Abt Books, Cambridge, MA.

Jöreskog, K. G. and Sörbom, D. (1986). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods*. Scientific Software Inc., Mooresville IN.

Juessen, F. (2009). A Distribution Dynamics Ato regional GDP Convergence in Unified Germany. *Empirical Economics*, 37(3):627–652.

Kaiser, H. (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.

Kapetanios, G. (2010). A Testing Procedure for Determining the Number of Factors in Approximate Factor Models With Large Datasets. *Journal of Business and Economic Statistics*, 28(3):397–409.

Karr, W. (1983). Aspekte saisonaler Arbeitslosigkeit. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, 16(1):17–27.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kaufman, G. M. and Press, S. J. (1973). Bayesian Factor Analysis. Working Papers 662–673, Massachusetts Institute of Technology (MIT), Sloan School of Management.

Kaufmann, S. and Schumacher, C. (2012). Finding Relevant Variables in Sparse Bayesian Factor Models: Economic Applications and Simulation Results. Discussion Papers 29/2012, Deutsche Bundesbank, Research Centre.

Kaufmann, S. and Schumacher, C. (2013). Bayesian Estimation of Sparse Dynamic Factor Models With Order-Independent Identification. Working Paper 1304, Study Center Gerzensee.

Kloek, T. and van Dijk, H. K. (1978). Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo. *Econometrica*, 46(1):1–19.

Koop, G. (2003). *Bayesian Econometrics*. John Wiley & Sons.

Koop, G. and Korobilis, D. (2010). Bayesian Multivariate Time Series Methods for Empirical Macroeconomics. *Foundations and Trends(R) in Econometrics*, 3(4):267–358.

Koopmans, T. (1947). Measurement Without Theory. *The Review of Economics and Statistics*, 29(3):161–172.

Korobilis, D. and Gilmartin, M. (2011). The Dynamic Effects of U.S. Monetary Policy on State Unemployment. Working Paper Series 12-11, The Rimini Centre for Economic Analysis.

Koschat, M. and Swayne, D. (1991). A Weighted Procrustes Criterion. *Psychometrika*, 56(2):229–239.

Kose, M. A., Otrok, C., and Prasad, E. (2012). Global Business Cycles: Convergence or Decoupling? *International Economic Review*, 53(2):511–538.

Kose, M. A., Otrok, C., and Prasad, E. S. (2008). Global Business Cycles: Convergence or Decoupling? NBER Working Papers 14292, National Bureau of Economic Research, Inc.

Kose, M. A., Otrok, C., and Whiteman, C. H. (2003). International Business Cycles: World, Region, and Country-Specific Factors. *American Economic Review*, 93(4):1216–1239.

Kristof, W. (1964). Die beste orthogonale Tranformation zur gegenseitigen Überführung zweier Factorenmatrizen. *Diagnostica*, 10:87–90.

Krolzig, H.-M. (2003). General-to-Specific Model Selection Procedures for Structural Vector Autoregressions. Economics Papers 2003-W15, Economics Group, Nuffield College, University of Oxford.

Kydland, F. E. and Prescott, E. C. (1982). Time to Build and Aggregate Fluctuations. *Econometrica*, 50(6):1345–1370.

Ladiray, D. and Quenneville, B. (2001). *Seasonal Adjustment with the X-11 Method*. Springer, New York.

Lawley, D. N. (1940). The Estimation of Factor Loadings by the Method of Maximum Likelihood. *Proceedings of the Royal Society of Edinburgh*, 60:64–82.

Ledermann, W. (1937). On the Rank of the Reduced Correlational Matrix in Multiple-Factor Analysis. *Psychometrika*, 2(2):85–93.

Lee, S.-Y. (1981). A Bayesian Approach to Confirmatory Factor Analysis. *Psychometrika*, 46(2).

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach*. John Wiley & Sons, Chichester.

Li, Q. and Lin, N. (2010). The Bayesian Elastic Net. *Bayesian Statistics*, 5(1):151–170.

Lissitz, R., Schönemann, P., and Lingoes, J. (1976). A Solution to the Weighted Procrustes Problem in Which the Transformation is in Agreement With the Loss Function. *Psychometrika*, 41(4):547–550.

Loken, E. (2005). Identification Constraints and Inference in Factor Analysis Models. *Structural Equation Modeling*, 12:232–244.

Lopes, H. F. and West, M. (2004). Bayesian Model Assessment in Factor Analysis. *Statistica Sinica*, 14:41–67.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). Sparse Statistical Modelling in Gene Expression Genomics. In Do, K. A., Mueller, P., and Vannucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, Cambridge UK.

Lucas, R. (1976). Econometric Policy Evaluation: A Critique. *Carnegie-Rochester Conference Series on Public Policy*, 1(1):19–46.

Lucas, R. E. (1977). Understanding Business Cycles. *Carnegie-Rochester Conference Series on Public Policy*, 5(1):7–29.

Lütkepohl, H. (2014). Structural Vector Autoregressive Analysis in a Data Rich Environment: A Survey. Discussion Papers of DIW Berlin 1351, DIW Berlin, German Institute for Economic Research.

Ma, H. and Zhao, H. (2013). Application of Bayesian Sparse Factor Analysis Models in Bioinformatics. In Do, K.-A., Qin, Z. S., and Vannucci, M., editors, *Advances in Statistical Bioinformatics*, chapter 17, pages 350–365. Cambridge University Press.

Malsiner-Walli, G. and Wagner, H. (2011). Comparing Spike and Slab Priors for Bayesian Variable Selection. *Austrian Journal of Statistics*, 40(4):241–264.

Marsh, H. W., Balla, J. R., and McDonald, R. P. (1988). Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size. *Psychological Bulletin*, 103.

Mehra, R. H. (1974). Topics in Stochastic Control Theory Identification in Control and Econometrics: Similarities and Differences. In *Annals of Economic and Social Measurement, Volume 3, number 1*, NBER Chapters, pages 21–48. National Bureau of Economic Research, Inc.

Millsap, R. E. (2001). When Trivial Constraints Are Not Trivial: The Choice of Uniqueness Constraints in Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(1):1–17.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83:1023–1032.

Moench, E., Ng, S., and Potter, S. (2009). Dynamic Hierarchical Factor Models. Staff Reports 412, Federal Reserve Bank of New York.

Molenaar, P. (1985). A Dynamic Factor Model for the Analysis of Multivariate Time Series. *Psychometrika*, 50(2):181–202.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

Mulaik, S. A. (2010). *Foundations of Factor Analysis*. Chapman & Hall, Boca Raton, FL, 2nd edition.

Muth, J. F. (1961). Rational Expectations and the Theory of Price Movements. *Econometrica*, 29(6):315–335.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313.

Neuhaus, J. O. and Wrigley, C. (1954). The Quartimax Method. *British Journal of Statistical Psychology*, 7(2):81–91.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

Ni, S. and Sun, D. (2005). Bayesian estimates for vector autoregressive models. *Journal of Business & Economic Statistics*, 23:105–117.

Onatski, A. (2010). Determining the Number of Factors from Empirical Distribution of Eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Otrok, C. and Whiteman, C. H. (1998). Bayesian Leading Indicators: Measuring and Predicting Economic Conditions in Iowa. *International Economic Review*, 39(4):997–1014.

Owyang, M. T., Piger, J., and Wall, H. J. (2005). Business Cycle Phases in U.S. States. *The Review of Economics and Statistics*, 87(4):604–616.

Owyang, M. T., Piger, J., and Wall, H. J. (2013). Discordant City Employment Cycles. *Regional Science and Urban Economics*, 43(2):367–384.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103:681–686.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(6):559–572.

Press, S. J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart and Winston.

Press, S. J. and Shigemasu, K. (1989). Bayesian Inference in Factor Analysis. In Press, S., Gleser, L., Perlman, M., and Sampson, A., editors, *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, pages 271–287. Springer Verlag.

Press, S. J. and Shigemasu, K. (1997). Bayesian Inference in Factor Analysis - Revised. Department of Statistics - Technical Report 243, University of California Riverside.

Quah, D. and Sargent, T. J. (1993). A Dynamic Index Model for Large Cross Sections. In *Business Cycles, Indicators and Forecasting*, NBER Chapters, pages 285–310. National Bureau of Economic Research, Inc.

Redner, R. A. and Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, 26(2):195–239.

Reinsel, G. (1983). Some Results on Multivariate Autoregressive Index Models. *Biometrika*, 70(1):145–156.

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures With an Unknown Number of Components. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 59(4):731–792.

Robert, C. P. (2014). Bayesian Computational Tools. *Annual Review of Statistics and Its Application*, 1(1):153–177.

Roppert, J. and Fischer, G. (1965). *Lineare Strukturen in Mathematik und Statistik unter besonderer Berücksichtigung der Faktoren- und Transformationsanalyse*. Number 1 in Arbeiten aus dem Institut für Höhere Studien und Wissenschaftliche Forschung. Physica, Vienna.

Rotemberg, J. and Woodford, M. (1997). An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy. In *NBER Macroeconomics Annual 1997, Volume 12*, NBER Chapters, pages 297–361. National Bureau of Economic Research, Inc.

Rubin, D. and Thayer, D. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Rubin, D. and Thayer, D. (1983). More on EM for ML factor analysis. *Psychometrika*, 48(2):253–257.

Sargent, T. J. and Sims, C. A. (1977). Business Cycle Modeling Without Pretending to Have Too Much a Priori Economic Theory. Working Papers 55, Federal Reserve Bank of Minneapolis.

Schanne, N., Wapler, R., and Weyh, A. (2010). Regional Unemployment Forecasts with Spatial Interdependencies. *International Journal of Forecasting*, 26(4):908–926.

Scheffé, H. (1953). A Method for Judging all Contrasts in the Analysis of Variance. *Biometrika*, 40(1):87–104.

Scherrer, W. and Deistler, M. (1998). A Structure Theory for Linear Dynamic Errors-in-Variables Models. *SIAM Journal on Control and Optimization*, 36(6):2148–2175.

Schönemann, P. H. (1966). A Generalized Solution to the Orthogonal Procrustes Problem. *Psychometrika*, 31(1):1–10.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Shiller, R. J. (1978). Rational Expectations and the Dynamic Structure of Macroeconomic Models : A Critical Review. *Journal of Monetary Economics*, 4(1):1–44.

Shumway, R. and Stoffer, D. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1):1–48.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL.

Smith, J. O. (2007). *Introduction to Digital Filters with Audio Applications*. W3K Publishing, http://www.w3k.org/books.

Song, X.-Y. and Lee, S.-Y. (2012). *Basic and Advanced Bayesian Structural Equation Modeling*. John Wiley & Sons, Chichester.

Spearman, C. (1904). "General Intelligence" - Objectively Determined and Measured. *American Journal of Psychology*, 15:201–293.

Sperrin, M., Jaki, T., and Wit, E. (2010). Probabilistic Relabelling Strategies for the Label Switching Problem in Bayesian Mixture Models. *Statistics and Computing*, 20(3):357–366.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stephens, M. (2000). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):795–809.

Stock, J. H. and Watson, M. W. (1989). New Indexes of Coincident and Leading Economic Indicators. In *NBER Macroeconomics Annual 1989, Volume 4*, NBER Chapters, pages 351–409. National Bureau of Economic Research, Inc.

Stock, J. H. and Watson, M. W. (1998). Diffusion Indexes. NBER Working Papers 6702, National Bureau of Economic Research, Inc.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335.

Stock, J. H. and Watson, M. W. (2002a). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2012). Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Tanner, M. A. and Wong, W. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82:528–540.

Ten Berge, J. M. F. and Sočan, G. (2007). The set of feasible solutions for reliability and factor analysis. In Lee, S.-Y., editor, *Handbook of Latent Variable and Related Models*. Elsevier, Burlington, MA.

Thurstone, L. L. (1931). Multiple Factor Analysis. *Psychological Review*, 38:406–427.

Thurstone, L. L. (1935). *The Vectors of Mind*. The University of Chicago Press, Chicago.

Thurstone, L. L. (1938). A New Rotational Method in Factor Analysis. *Psychometrika*, 3(4):199–218.

Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tinbergen, J. (1932). Ein Problem der Dynamik. *Zeitschrift für Nationalökonomie*, 3:169–184.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S. (2003). Ensemble Square Root Filters. *Monthly Weather Review*, 131:1485–1490.

Titsias, M. and Lázaro-Gredilla, M. (2011). Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira,

F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 2339–2347. NIPS.

Watson, M. W. and Engle, R. F. (1983). Alternative Algorithms for the Estimation of Dynamic Factor, MIMIC, and Varying Coefficient Regression Models. *Journal of Econometrics*, pages 385–400.

West, M. (2003). Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. In *Bayesian Statistics 7*, pages 723–732. Oxford University Press.

Zellner, A. (1970). Estimation of Regression Relationships Containing Unobservable Independent Variables. *International Economic Review*, 11(3):441–454.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15:265–286.

Kooperationen

Abgesehen von meiner Person haben Christian Aßmann und Jens Boysen-Hogrefe im Rahmen gemeinsamer Forschungsprojekte an Kapitel 3 der vorliegenden Arbeit mitgewirkt. Mein Anteil ist hierbei wie folgt gegeben:

Kapitel 3:

Mein Beitrag zu dieser Arbeit besteht in

- **Beobachtung des Rotationsproblems bei bayesianischer Schätzung von Faktormodellen**
- **Herleitung vollständig bedingter Dichten, des unrestringierten Gibbs-Samplers und der Regeln für orthogonale Transformationen**
- **Wahl der quadratischen Verlustfunktion (ausformuliert von Christian Aßmann)**
- **Darstellung des Zusammenhangs mit dem (gewichteten) orthogonalen Procrustes-Problem und Matrixzerlegungstheorem für das dynamische Faktormodell**
- **Beweise zu Proposition 3.3.1 (teilweise) und Proposition 3.3.2 (weitestgehend)**
- **Programmierung des unrestringierten Samplers und des WOP-Algorithmus für Illustration, Simulationsstudie und Anwendung**

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich, abgesehen von den auf der vorangegangenen Seite präzisierten Kooperationen mit Christian Aßmann und Jens Boysen-Hogrefe, meine Doktorarbeit "Model Identification in Bayesian Analysis of Static and Dynamic Factor Models" selbständig und ohne fremde Hilfe angefertigt habe und dass ich alle von anderen Autoren wörtlich übernommenen Stellen, wie auch die sich an die Gedanken anderer Autoren eng anlehnenden Ausführungen meiner Arbeit, besonders gekennzeichnet und die Quellen nach den mir angegebenen Richtlinien zitiert habe.

Kiel, den 07. Januar 2015 Markus Pape