Bayesian Factor Models for Clustering and Spatiotemporal Analysis

Hwasoo Shin

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Marco A. R. Ferreira, Co-chair
Allison N. Tegge, Co-chair
Christopher T. Franck
Inyoung Kim

May 14th, 2024
Blacksburg, Virginia

Keywords: Bayesian Factors Model, Spatiotemporal data, Clustering methods, Dimension reduction

Bayesian Factor Models for Clustering and Spatiotemporal Analysis

Hwasoo Shin

ABSTRACT

Multivariate data is prevalent in modern applications, yet it often presents significant analytical challenges. Factor models can offer an effective tool to address issues associated with large-scale datasets. In this dissertation, we propose two novel Bayesian factors models. These models are designed to effectively reduce the dimensionality of the data, as the number of latent factors is typically much smaller than that of the observation vectors. Therefore, our proposed models can achieve substantial dimension reduction. Our first model is for spatiotemporal areal data. In this case, the region of interest is divided into subregions, and at each time point there is one univariate observation per subregion. Our model writes the vector of observations at each time point in a factor model form as the product of a vector of factor loadings and a vector of common factors plus a vector of error. Our model assumes that the common factor evolve through time according to a dynamic linear model. To represent the spatial relationships among subregions, each column of the factor loadings matrix is assigned an intrinsic conditional autoregressive (ICAR) priors. Therefore, we call our approach the Dynamic ICAR Spatiotemporal Factor Models (DIFM). Our second model, Bayesian Clustering Factor Model (BCFM) assumes latent factors and clusters are present in the data. We apply Gaussian mixture models on common factors to discover clusters. For both models, we develop MCMC to explore the posterior distribution of the parameters. To select the number of factors and, in the case of clustering methods, the number of clusters, we develop model selection criteria that utilize the Laplace-Metropolis estimator of the predictive density and BIC with integrated likelihood.

Bayesian Factor Models for Clustering and Spatiotemporal Analysis

Hwasoo Shin

GENERAL AUDIENCE ABSTRACT

Understanding large-scale datasets has emerged as one of the most significant challenges for researchers recently. This is particularly true for datasets that are inherently complex and nontrivial to analyze. In this dissertation, we present two novel classes of Bayesian factor models for two classes of complex datasets. Frequently, the number of factors is much smaller than the number of variables, and therefore factor models can be an effective approach to handle multivariate datasets. First, we develop Dynamic ICAR Spatiotemporal Factor Model (DIFM) for datasets collected on a partition of a spatial domain of interest over time. The DIFM accounts for the spatiotemporal correlation and provides predictions of future trends. Second, we develop Bayesian Clustering Factor Model (BCFM) for multivariate data that cluster in a space of dimension lower than the dimension of the vector of observations. BCFM enables researchers to identify different characteristics of the subgroups, offering valuable insights into their underlying structure.

Acknowledgments

I appreciate my advisor, Marco Ferreira, for his invaluable guidance through this Ph.D. journey at Virginia Tech. I am immensely grateful for mentorship, which has been instrumental in exploring the numerous approaches to research. I also thank you for your patience, insightful feedback, and the resources you generously provided. My research in Bayesian factor models was pleasant because I could work with you.

I give many thanks to Allison Tegge for her generous consideration in offering me the opportunity to work as a research assistant. I enjoyed being on a team to join a project on biostatistics studies, which I was particularly interested in. It was a delight to experience analyzing clinical trial datasets and learn the approaches we use in this field. Your guidance provided invaluable insights into the practical applications of statistics in the real world.

To my office mates, Tsering Dolkar, Steve Walsh, Jake Williams, and Shuangshuang Xu, I extend my gratitude for being a good friend. Your assistance, no matter how small, was encouraging to me. The feedback and suggestions you provided became invaluable strengths in writing an engaging paper. I am glad to study and work with such wonderful colleagues.

My family has always been my biggest fan. Without your support, I wouldn't have been persevere through the challenges of graduate school. To my parents, Kangho Shin and Kyonghee Han, you have been my role models, inspiring me to pursue my dreams of obtaining a doctoral degree at Virginia Tech. To my grandmother, Kisook Kim, thank you for your care and love even when I was far away from home. And to my sister, Gabrielle, you have been the world's best little sister, and you always will be. Your belief in me motivated my determination to succeed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we explore the background and the idea of this dissertation. Our research focuses on dimension reduction unsupervised methods, employing Bayesian approach to multivariate datasets. Bayesian Factor model is an effective method to offer solutions to issues with multidimensional datasets. Our first proposed model considers the spatiotemporal correlations of the observations. In Second, we develop a novel factor model to discover clusters within datasets. There are major concepts of this dissertation we will discuss in the following sections: unsupervised methods, factor models, spatiotemporal datasets and clustering models. Lastly, we provide the outline of this dissertation.

1.1 Unsupervised Methods

Unsupervised methods are widely used in machine learning to explore and understand the dataset. The goal is to discover patterns and relationships of the variables without guidance. Unlike supervised methods, unsupervised methods do not have a target output. Therefore, traditional train-test sets split and evaluation is inapplicable. Instead, unsupervised method detects the inherent structure embedded in the data.

Examples of unsupervised learning are visualization, association rules, clustering, and dimension reduction. Clustering methods are especially used to find the relationships of the sample. For example, a researcher can apply customer segmentation and market basket

analysis to define the subgroups of the customers, predict purchasing patterns, and apply personalized advertisements to attract customers. In journalism and politics, employing semantic and document clustering is helpful for categorizing articles due to opinion polarization. In genomics, this approach is also used to cluster RNA and DNA sequences based on similarities in their genetic makeup. Unsupervised clustering assigns the same subgroups to observations with similar characteristics, while different observations are grouped into distinct clusters. Unsupervised clustering helps us to understand the relationship and characteristics of the subjects. The clusters identified through the unsupervised method provide researchers with an improved strategy for approaching individual subjects. While supervised learning requires information on labels defined prior to the analysis, unsupervised method can be applied to unlabeled data. Supervised methods determine boundaries to separate observations, whereas unsupervised methods group observations based on their distances.

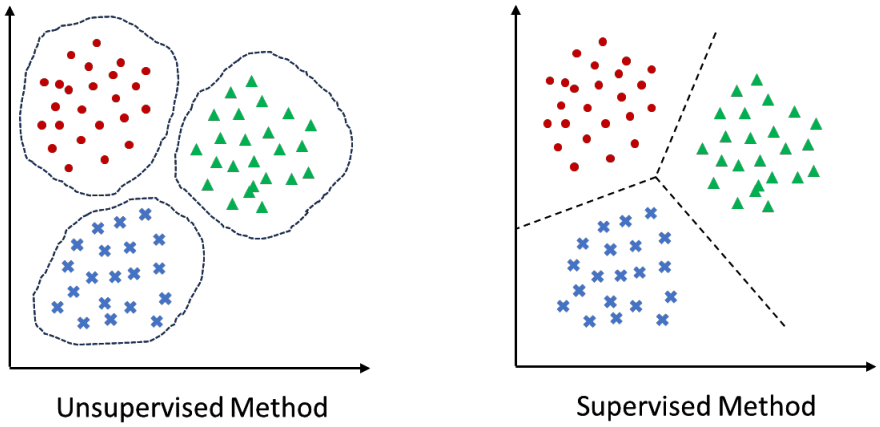


Figure 1.1: Supervised and Unsupervised Clustering

Unsupervised methods serve as effective tools for exploring and gaining insights into the associations within data, playing a crucial role in various machine learning advancements. Additionally, they offer cost savings of computations by eliminating the need for labeled data, unveil hidden patterns, and are feasible for handling large-scale datasets.

1.2 Bayesian Factor Models

Datasets with multiple variables pose a challenge for statistical analysis. Many problems require extracting valuable information from correlated variables and sparse datasets. In such cases, factor models emerge as a powerful strategy. These models identify factors as linear or non-linear combinations of variables, revealing primary trends for a more interpretable understanding of the data.

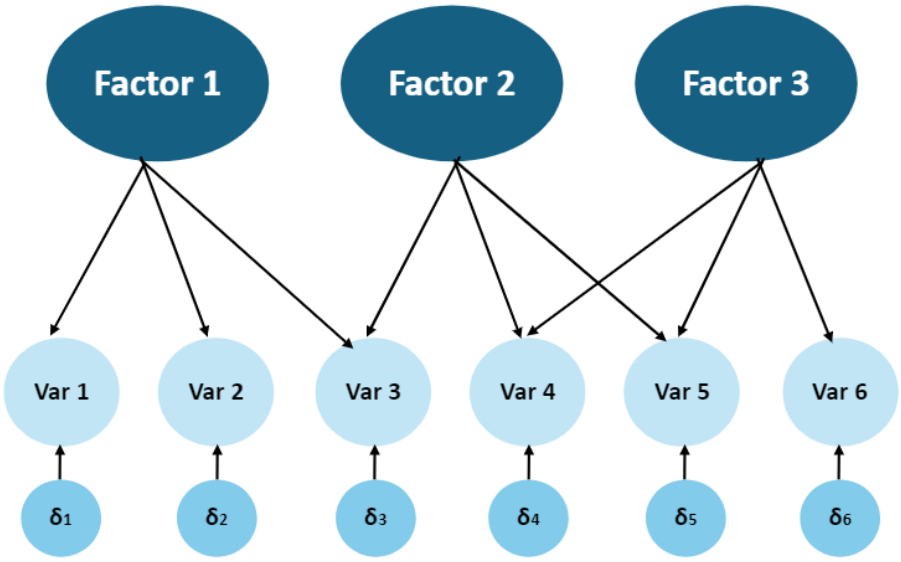


Figure 1.2: Factor Model Diagram

Researchers can streamline their analysis by focusing on a select few factors instead of utilizing the entire set of variables. This proves especially beneficial when dealing with datasets with numerous correlated independent variables that lack substantial information. Factor analysis results in a considerably smaller number of factors compared to the data dimension. The integration of the Bayesian approach with factor models enhances flexibility, allowing for the incorporation of prior distributions on the parameters. With this approach, Bayesian inference may explain variability and correlations of the parameters more efficiently compared to models without the approach. Factor models are pragmatic as they enable researchers to circumvent the need to incorporate every variable for analysis. Thus, they

decrease time and computation costs. Bayesian factor models are flexible to extend by adding structures to account for specific cases of dataset. For example, Bayesian factor models can address spatially correlated census data [25] or multivariate codependence of categorical data [56]. Bayesian factor models often require specifying the number of factors to initiate [37] the simulation. In this case, we have to carefully choose this value to begin with and compare the performance of the models across a range of different number of factors. On the other hand, some methods apply overfitting strategy that adopts infinite factors and reduces the size to automatically detect an appropriate number of factors [7]. While Bayesian factor models have numerous advantages, they often suffer from identifiability issues. While these issues do not typically affect predictions, they can significantly complicate interpretations. To prevent identifiability problems [45], we impose constraints on the model [23] or adjust the MCMC algorithm. Chapter 2 and Chapter 3 apply Bayesian factor models with additional structures described in Section 1.3 and Section 1.4, respectively.

1.3 Spatiotemporal Datasets

Observations in many studies are collected in several geographical regions over a series of time periods. This is common in various fields including crime analysis [49], disease mapping [39], ecological modeling [12], and numerous other research domains. We should incorporate an appropriate framework into the model to account for inherent correlations present in spatial data. In geostatistical analysis, selecting an appropriate variogram structure is crucial for modeling spatial dependencies in the data [20]. For areal data, we develop models that capture the relationships between neighboring subregions [43]. Subjects recorded over time are usually influenced by their previous time point measurements. Time series analysis accounts for dynamic behavior of observations and separates meaningful trends from noise. We compute the size of the autocorrelation and utilize it to derive predictions [16]. In Chapter 2, we adopt the intrinsic autoregressive prior (ICAR) for spatial correlations and

dynamic linear models for temporal structures.

1.4 Cluster Analysis

Cluster analysis has been widely practiced in statistics and computer science. By identifying latent groups within a dataset, clustering models give researchers insights into the underlying properties of observations. One common application is hierarchical clustering [13], which defines the clusters recursively according to the closest distances of observations. Another well-known method is centroid clustering [31], such as k-means. This method assigns observations to groups according to the initial centroids and iteratively adjusts both the centroids and cluster assignments throughout the process. Labels can be assigned to clusters to explain their characteristics, either based on the analysis itself or information in the current dataset. In Bayesian inference, one approach to clustering involves determining clusters with fixed numbers and subsequently comparing their performance [54]. Alternatively, we can use the multiplicative gamma process to fit an infinite cluster model [40] and narrow down the number of clusters during the MCMC. In our proposed Bayesian Clustering Factor Models outlined in Chapter 3, we specify both the number of clusters and factors and evaluate different settings to select a model with the optimal performance.

1.5 Dissertation Outline

The remainder of this dissertation is organized as follows. We propose two Bayesian factor models and possible extensions of this research. In Chapter 2, we introduce the Dynamic ICAR Spatiotemporal Factor Models (DIFM). DIFM considers the observations that evolve through time and spatial dependence from the adjacent areas. DIFM applies a dynamic linear model to the common factors and assumes ICAR priors to the factor loadings to account

for the spatial correlations. Chapter 3 describes the framework of the Bayesian Clustering Factor Models (BCFM). BCFM is a Bayesian factor model with a Gaussian mixture model on the common factors to identify clusters. In Chapter 4, we describe two R packages related to our papers that run models, assessments, and plots for the models proposed in Chapter 2 and Chapter 3. Finally, in Chapter 5, we give conclusions and discuss the extensions of the research.

Chapter 2

Dynamic ICAR Spatiotemporal Factors Model

2.1 Introduction

Factor models are one of the most widely used tools for the analysis of multivariate data. Factor models often achieve substantial data reduction by writing the often high dimensional vector of observations as a linear function of relatively few common factors. In particular, there has been growing research interest in Bayesian factor models [1, 2, 11, 23, 35, 37, 38, 57]. This growing interest is because Bayesian factor models allow the incorporation of important structures for the analysis of highly structured datasets, such as time series, spatial, and spatiotemporal data. In particular, Lopes et al. [37] and Lopes et al. [35] have proposed Bayesian factor models for spatiotemporal point-referenced data. However, many spatiotemporal datasets are for areal data. Here, we propose novel Bayesian factor models for spatiotemporal areal data.

The class of Bayesian spatiotemporal models for univariate areal data that we propose is a dynamic factor model for the vector of areal data observed at each time point. As such, our proposed model has spatial factor loadings and temporal common factors. Specifically, we assume that the common factors evolve through time according to dynamic linear models (DLM) [47, 58]. The use of DLMs allows the inclusion of important temporal structures, such as seasonality and different forms of trend. In addition, we assume that each column of the matrix of factor loadings follows an intrinsic conditional autoregressive (ICAR) model [5, 28, 29]. These ICAR models allow borrowing of information among spatial neighboring regions for the estimation of the factor loadings. We call our class of models the Dynamic

ICAR Spatiotemporal Factor Models (DIFM).

We develop methods for estimation and model selection for DIFMs. Specifically, we develop an efficient Markov Chain Monte Carlo (MCMC) algorithm [21] for the exploration of the posterior distribution. The latent process associated with the common factors is simulated using a Forward Filter Backward Sampler (FFBS) algorithm [10, 19]. Importantly, the elements of the matrix of factor loadings are simulated in one single step from their joint full conditional distribution. For the other parameters in the model, we assume conjugate priors which imply full conditional distributions that are straightforward to sample from. Application of this algorithm to a simulated dataset shows that the algorithm works well at estimating the parameters. Finally, for model selection for DIFMs, we develop a Laplace-Metropolis approximation to the predictive density that we call the Laplace-Metropolis predictive density. As shown in Section 2.5.1, the Laplace-Metropolis predictive density can be used to successfully select the number of factors for DIFMs.

We illustrate the use of our framework with an application to the *VSRR Provisional Drug Overdose Death Counts* (PDODC) dataset from the *Centers for Disease Control and Prevention* (CDC). Specifically, we analyze the monthly number of deaths per 100,000 people by drug overdose in each of the 48 contiguous states of the United States from January 2015 to February 2021. Section 2.2 provides an exploratory data analysis of this dataset that motivates the development of DIFMs. In addition, Section 2.5.2 provides an analysis of this dataset using DIFMs. Compared with the spatiotemporal CAR ANOVA model [30] implemented in the `CARBayesST` package [32], our proposed DIFMs has much more favorable predictive performance. Finally, the DIFM analysis provides meaningful results: The estimated factor loadings exhibit interesting spatial patterns, and the estimated common factors shed light on the impact of the COVID pandemic on the drug overdose epidemic in the United States.

The remainder of the chapter is organized as follows. In Section 2.2, we perform an exploratory data analysis of the PDODC dataset using principal components. In Section 2.3, we introduce the class of Bayesian dynamic ICAR factor models that we propose. In Section 2.4, we develop a Gibbs sampler for the estimation of the parameters and we propose a Laplace-Metropolis predictive density for model selection. In Section 2.5, we present applications to both a simulated dataset and a real dataset. Section 2.6 concludes with a discussion

and possible future directions.

2.2 Data Exploration

We present in this section an exploratory data analysis of the monthly number of deaths by drug overdose in each of the 48 contiguous states of the United States from January 2015 to February 2021. These data are publicly available from the *VSRR Provisional Drug Overdose Death Counts* dataset from the *Centers for Disease Control and Prevention* at https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm. To normalize and variance-stabilize the data, we consider the square root of the monthly standardized number of deaths per 100,000 inhabitants.

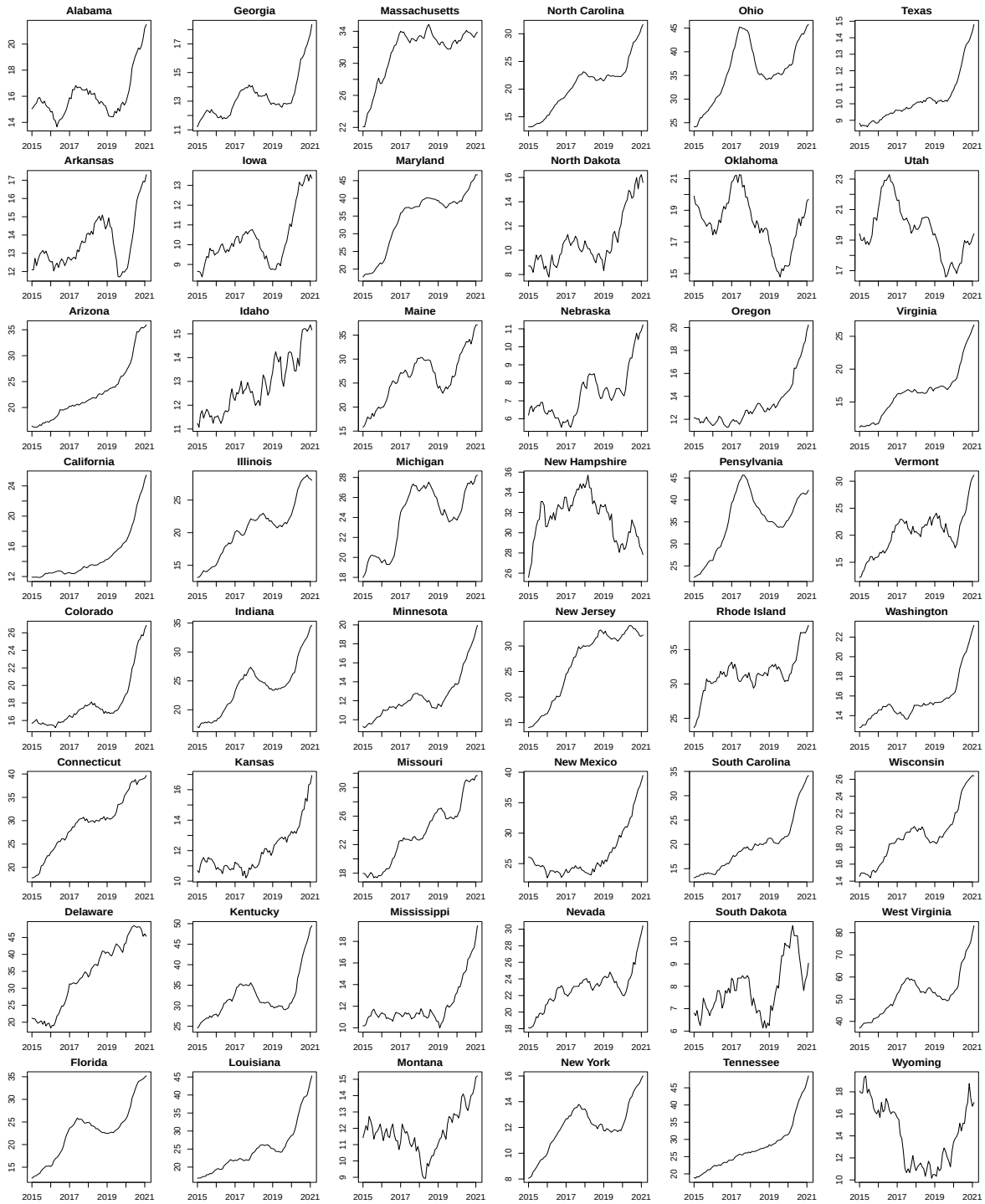


Figure 2.1: Figure 1: Monthly drug overdose death counts per 100,000 inhabitants by state of the contiguous United States from January 2015 to February 2021.

Figure 2.1 shows monthly drug overdose death counts per 100,000 inhabitants by state from January 2015 to February 2021. The mortality rate due to drug overdose increased from 2015

to 2020 in most states, except for a few states such as Utah and Wyoming. In addition, there is a difference in how the mortality rate increased. Some states experienced a steady rise in the mortality rate through time, such as Arizona, California and Tennessee. In another pattern, some states saw a decrease between around 2018 and 2019, as for example, Alabama, Florida and Georgia. Finally, most states had a substantial increase in the mortality rate due to drug overdose starting at the beginning of the COVID pandemic in the United States in March 2020.

To implement a DIFM, we first perform an exploratory principal component analysis (PCA) to help us make key decisions for the factor model. Figure 2 shows the scree plot for the PDODC dataset. In this dataset, the first principal component explains about 72.5% of the variability, the second explains 12% of the variability, the third explains around 6.2%, and the fourth explains 3.5% of the variability. Thus, the four first principal components explain a total of 94.2% of the variability. Hence, the PCA suggests an initial number of 4 factors for our factor model. While we first explored a DIFM with 4 factors, Section 2.4.2 proposes a more formal Bayesian model selection approach to decide on the optimal number of factors for our DIFM.

We note that, as explained in Section 2.3, Bayesian factor models use a hierarchical structural constraint to ensure identifiability. This constraint assumes a lower triangular matrix of factor loadings, which makes the analysis dependent on the order of the variables. Hence, we use the exploratory PCA to decide on the order of the variables, which correspond here to the order of the states. We choose the order of the states according to the magnitude of the loadings in the different principal components. In addition, we choose the order so that the states in the first positions are located far away from each other. Specifically, in the first principal component the state with the largest loading is Virginia. Thus, we choose Virginia to be the first state in our DIFM. In addition, in the second principal component Montana has one of the largest loadings and is distant from Virginia; thus Montana becomes the second state in our DIFM. Following similar considerations, Iowa and Alabama are chosen to be the third and fourth states in our DIFM.

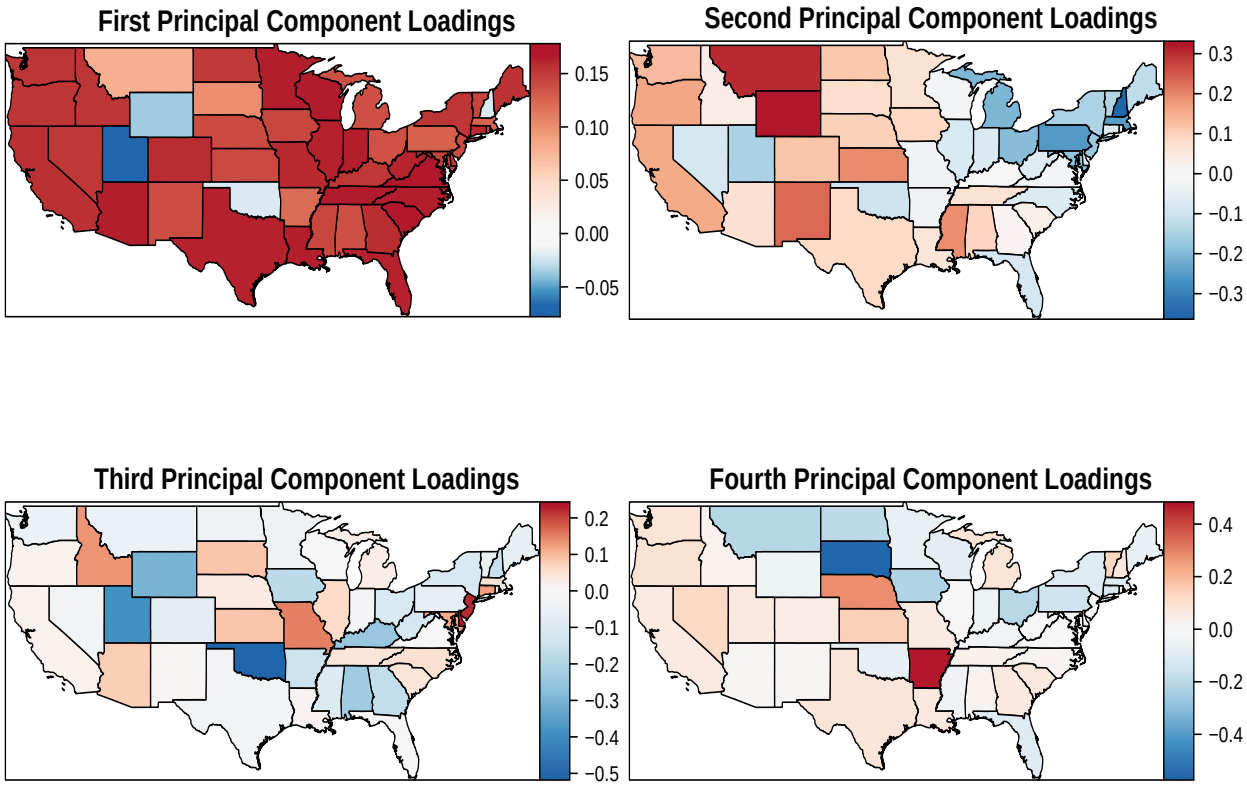


Figure 2.2: Figure 3: Maps of factor loadings for the first four principal components.

Figure 2.2 presents the maps of factor loadings for the first four principal components. In these maps, shades of red represent positive values, shades of blue represent negative values, and white represents values close to zero. The first principal component has most of its loadings positive with values above 0.1. There are only 2 states, Utah and Wyoming, that have blue color. The second principal component has many states in the east in blue and many states in the west in shades of red, indicating presence of spatial correlation. The third principal component has most of the states in light color, with some states on the East Coast and some states in the Midwest having positive loadings. Finally, the fourth principal component has states in the West Coast and in the south in light orange, whereas states in the north appear in light blue, indicating presence of spatial dependence.

In summary, from the maps of the loadings for the different principal components, it seems reasonable to assume that the loadings are spatially correlated. This motivates the class of

DIFMs that we present in the next section.

2.3 Dynamic ICAR Spatiotemporal Factor Models

In this section, we present our DIFMs for spatiotemporal areal data. We assume that the region of study is partitioned into r subregions. In our motivating example, the subregions are the states in the contiguous United States. The variable of interest in each subregion is observed at n time points. Let $\mathbf{y}_t$ be the r-dimensional vector of observations at time t $(t = 1, 2, \ldots, n.)$

We assume that the spatiotemporal behavior of the r subregions can be represented by k factors, where usually k is much smaller than r. Specifically, we assume the model

$$\mathbf{y}_t = \mathbf{B}\mathbf{x}_t + \mathbf{v}_t, \tag{2.1}$$

where $\mathbf{x}_t$ is the k-dimensional vector of factors at time t, $\mathbf{B}$ is an $r \times k$ matrix of factor loadings, and $\mathbf{v}_t$ is the r-dimensional vector of errors at time t. We assume that the observational error vector $\mathbf{v}_t$, $t = 1, 2, \ldots, n$ is independent over time and follows a Gaussian distribution $\mathbf{v}_t \sim \mathrm{N}(0, \mathbf{V})$, where $\mathbf{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_r^2)$. Each of the variances $\sigma_1^2, \ldots, \sigma_r^2$ is specific to one of the r subregions, and thus they are known as idiosyncratic variances.

We assume that the vector of factors $\mathbf{x}_t$ follows a dynamic linear model [47, 58]. Specifically, we assume the general model

$$\mathbf{x}_t = \mathbf{F}\boldsymbol{\theta}_t, \tag{2.2}$$

$$\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \boldsymbol{\omega}_t \sim \mathrm{N}(0, \mathbf{W}), \tag{2.3}$$

where $\boldsymbol{\theta}_t$ is a latent process that allows great flexibility in the description of the temporal evolution of $\mathbf{x}_t$. Specifically, $\boldsymbol{\theta}_t$ may encode different types of temporal trends as well as seasonality. For example, in our application we assume a second-order polynomial DLM and specify θ_t as a vector of dimension $2k$ that contains the level and the gradient of $\mathbf{x}_t$ at time t. In addition, the evolution matrix $\mathbf{G}$ describes the temporal evolution of the latent process $\boldsymbol{\theta}_t$. Further, $\boldsymbol{\omega}_t$ is a $2k$-dimensional innovation vector with a dense covariance matrix $\mathbf{W}$.

Finally, the matrix $\mathbf{F}$ relates the vector of common factors $\mathbf{x}_t$ to the appropriate elements of the latent process $\boldsymbol{\theta}_t$.

In the case of the second-order polynomial DLM that we consider, $\boldsymbol{\theta}_t = (\theta_{t,1}, \theta_{t,2}, \ldots, \theta_{t,2k})^T$ is a vector of dimension $2k$ where $(\theta_{t,1}, \theta_{t,3}, \ldots, \theta_{t,2k-1})^T$ and $(\theta_{t,2}, \theta_{t,4}, \ldots, \theta_{t,2k})^T$ are respectively the level and the gradient of the vector of common factors $\mathbf{x}_t$. Thus, the matrix $\mathbf{F}$ that relates $\mathbf{x}_t$ to $\boldsymbol{\theta}_t$ is a $k \times 2k$ matrix of the form

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{bmatrix}.$$

The evolution matrix $\mathbf{G}$ has dimension $2k \times 2k$ and satisfies $\theta_{t,2j-1} = \theta_{t-1,2j-1} + \theta_{t-1,2j} + \omega_{t,2j-1}$ and $\theta_{t,2j} = \theta_{t-1,2j} + \omega_{t,2j}$, $j = 1, \ldots, k$. Therefore, $\mathbf{G} = blockdiag(\mathbf{G}_0, \ldots, \mathbf{G}_0)$ where

$$\mathbf{G}_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The specification of the factor loadings matrix $\mathbf{B}$ is crucial in our dynamic ICAR factor model. An important point to consider is the need for constraints on the matrix $\mathbf{B}$ to ensure identifiability of the model. Specifically, for any invertible $k \times k$ matrix $\mathbf{A}$, substituting $\mathbf{B}$ and $\mathbf{x}_t$ in Equation (3.1) by, respectively, $\mathbf{B}^* = \mathbf{BA}$ and $\mathbf{x}_t^* = \mathbf{A}^{-1}\mathbf{x}_t$ would lead to the same model. To ensure identifiability, we impose a hierarchical structural constraint that assumes that $\mathbf{B}$ is a full-rank block lower triangular matrix with diagonal elements equal to

1 [2, 23, 38]. Specifically, we assume $\mathbf{B}$ has the form

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ b_{2,1} & 1 & 0 & \dots & 0 \\ b_{3,1} & b_{3,2} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{k,1} & b_{k,2} & b_{k,3} & \dots & 1 \\ b_{k+1,1} & b_{k+1,2} & b_{k+1,3} & \dots & b_{k+1,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{r,1} & b_{r,2} & b_{r,3} & \dots & b_{r,k} \end{bmatrix}.$$

To account for the spatial dependence among the factor loadings for neighboring subregions, we assume that each column of the matrix of factor loadings $\mathbf{B}$ follows an intrinsic conditional autoregressive model [5, 6, 28, 29]. Specifically, we assume for the jth column $\mathbf{B}_j, j = 1, \dots, k$, the density

$$p(\mathbf{B}_j) \propto \exp\left(-\frac{1}{2\tau_j}\mathbf{B}_j^T\mathbf{H}\mathbf{B}_j\right), \tag{2.4}$$

where $\mathbf{H}$ is a precision matrix that accounts for the spatial dependence among neighboring subregions and τ_j controls the strength of spatial correlation among factor loadings. Specifically, if subregions i and j are neighbors, then the corresponding element of the matrix $\mathbf{H}$ is $h_{ij} = -g_{ij}$ where g_{ij} measures the strength of the association between subregions i and j. If subregions i and j are not neighbors, then $g_{ij} = 0$. Finally, the ith diagonal element of matrix $\mathbf{H}$ is $h_{ii} = \sum_{j \neq i} g_{ij}$. For example, a widely used choice for $\mathbf{H}$ assumes $g_{ij} = 1$ if i and j share a border, and $g_{ij} = 0$ otherwise. In that case, h_{ii} is equal to the number of neighbors of subregion i. Further, we assume that there are no islands which implies that the matrix $\mathbf{H}$ has one eigenvalue equal to 0 and all other eigenvalues larger than zero. Note that we assume this prior for each column of $\mathbf{B}$. Let $\mathbf{B}^*_{\cdot j} = \mathbf{B}_{(j+1):r,j}$ be the jth column of $\mathbf{B}$ without the first j elements that are fixed. In addition, let $\mathbf{H}^*_j = \mathbf{H}_{(j+1):r,(j+1):r}$. Then, the conditional distribution of $\mathbf{B}^*_{\cdot j}$ given $\mathbf{B}_{1:j,j} = (0, \dots, 0, 1)^T$ is multivariate normal with mean vector $\mathbf{h}_j = -\mathbf{H}^{*-1}_j\mathbf{H}_{(j+1):r,j}$ and precision matrix $\mathbf{H}^*_j$. Finally, when we simulate $\mathbf{B}$ in the Gibbs sampler proposed in Section 2.4.1, we simulate the unknown values of $\mathbf{B}$ conditional

on the fixed values of $\mathbf{B}$ following the "sampling under a hard constraint" approach presented in pages 36 and 37 of Rue and Held [50]. We provide details of this approach specifically for DIFMs in Section 2.4.1.

2.3.1 Priors for initial states and hyperparameters

We complete the specification of the model with the assignment of prior distributions for the initial states and the hyperparameters. For simplicity of exposition, here we use the notation for DLMs from West and Harrison [58]. First, we assign a prior distribution for the latent process at time 0. Let D_0 be the prior information at time 0. Further, let $D_t = D_{t-1} \cup \{\mathbf{Y}_t\}$ be the information up to time t. Then, we assign for the initial state of the latent process $\boldsymbol{\theta}_0$ given D_0 a multivariate Gaussian distribution $\mathrm{N}(\mathbf{m}_0, \mathbf{C}_0)$.

Second, we assign the priors for the hyperparameters. For efficient computations, we assign conditional conjugate priors for the hyperparameters. Specifically, we assign for the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_r^2$ inverse gamma priors $IG(n_\sigma/2, n_\sigma s_\sigma^2/2)$. For the evolution covariance matrix $\mathbf{W}$ we assign an inverse Wishart prior $\mathbf{W} \sim IW(\mathbf{S}_W, n_W)$. For $\tau_1, \ldots, \tau_k$ we assign independent inverse gamma priors $\tau_j \sim IG(n_\tau/2, n_\tau s_\tau^2/2)$.

2.4 Statistical Inference

This section presents the statistical methods that we propose to analyze data using DIFMs. Section 2.4.1 presents an MCMC algorithm to explore the posterior distribution. Section 2.4.2 presents a model selection approach to decide among competing DIFMs.

2.4.1 Posterior exploration

This section presents a Gibbs sampler that facilitates inference for the parameters of DIFMs. We present below the full conditional distributions of the parameters of the model. In particular, we provide details about how to perform the nontrivial task of jointly sampling all the unconstrained elements of the matrix of factor loadings $\mathbf{B}$.

The full conditional distribution of each observational error variance $\sigma_j^2, j = 1, \ldots, r$, is the inverse gamma distribution

$$\sigma_j^2 | \mathbf{Y}, \mathbf{X} \sim IG\left(\frac{n_\sigma + n}{2}, \frac{n_\sigma s_\sigma^2 + \sum_{i=1}^n \mathbf{Y}_{ij} - \mathbf{X}_{i.}\mathbf{B}_{j.}^T}{2}\right). \tag{2.5}$$

The full conditional distribution of the evolution covariance matrix $\mathbf{W}$ is the inverse Wishart distribution

$$\mathbf{W} | \mathbf{Y}, \mathbf{X} \sim IW\left(\sum_{t=2}^{T} (\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})^T + \mathbf{S}_W, n_W + n - 1\right). \tag{2.6}$$

The parameter τ_j that controls the strength of spatial correlation among the factor loadings for the jth factor, $j = 1, \ldots, k$, has the inverse gamma full conditional distribution

$$\tau_j | \mathbf{X}, \mathbf{Y} \sim IG\left(\frac{n_\tau + r - j}{2}, \frac{n_\tau s_\tau^2 + (\mathbf{B}_{.j}^* - \mathbf{h}_j)^T \mathbf{H}_j^* (\mathbf{B}_{.j}^* - \mathbf{h}_j)}{2}\right). \tag{2.7}$$

The simulation of the matrix of factor loadings $\mathbf{B}$ has to take into account the hierarchical structural constraint and the spatial dependence among the factor loadings. Due to the hierarchical structural constraint, in the k-factor model the matrix of factor loadings $\mathbf{B}$ has $rk - k(k+1)/2$ free elements and the remaining $k(k+1)/2$ elements are fixed at 0 or 1. To account for the spatial dependence, we first obtain the joint full conditional distribution of all the elements of $\mathbf{B}$ and then we use standard multivariate Gaussian results to obtain the full conditional distribution of the free elements of $\mathbf{B}$ given the fixed elements of $\mathbf{B}$ and all the other quantities of the model.

Let us first consider the full conditional distribution of $\mathbf{b} = \text{vec}(\mathbf{B})$, the vector obtained by stacking the columns of the matrix $\mathbf{B}$. Let $\mathbf{H}_B = \text{diag}(1/\tau_1, \ldots, 1/\tau_k) \otimes \mathbf{H}$ and $\mathbf{U}_i = \mathbf{X}_{i.} \otimes \mathbf{I}_r$, where $\mathbf{I}_r$ is the $r \times r$ identity matrix and $\otimes$ is the Kronecker product. Then, the full conditional distribution of $\mathbf{b}$ is a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_B = \boldsymbol{\Sigma}_B \sum_{i=1}^n \mathbf{U}^T \mathbf{V}^{-1} \mathbf{y}_i$ and covariance matrix $\boldsymbol{\Sigma}_B = \left(\mathbf{H}_B + \sum_{i=1}^n \mathbf{U}_i^T \mathbf{V}^{-1} \mathbf{U}_i\right)^{-1}$.

Let $m = (1, r+1, r+2, \ldots, (k-1)r + 1, \ldots, (k-1)r + k)$ be the vector that contains the positions of the fixed elements of $\mathbf{B}$ in the vector $\mathbf{b}$, with corresponding fixed values

$\mathbf{c} = (1, 0, 1, 0, 0, \ldots, 0, \ldots, 1)^T$. Let $\mathbf{b}_m$ and $\mathbf{b}_{-m}$ be the subvectors of $\mathbf{b}$ corresponding to its fixed and free elements, respectively. Using analogous notation, let $\boldsymbol{\mu}_{B,m}$ and $\boldsymbol{\mu}_{B,-m}$ be the corresponding subvectors of $\boldsymbol{\mu}_B$. Similarly, let $\boldsymbol{\Sigma}_{B,m}$, $\boldsymbol{\Sigma}_{B,-m}$, $\boldsymbol{\Sigma}_{B,(m,-m)}$, and $\boldsymbol{\Sigma}_{B,(-m,m)}$ be the corresponding submatrices of the matrix $\boldsymbol{\Sigma}_B$. Then, the full conditional distribution of $\mathbf{b}_{-m}$ given $\mathbf{b}_m$ is a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_B^*$ and covariance matrix $\boldsymbol{\Sigma}_B^*$ where

$$\boldsymbol{\mu}_B^* = \boldsymbol{\mu}_{B,-m} + \boldsymbol{\Sigma}_{B(-m,m)}\boldsymbol{\Sigma}_{B(m,m)}^{-1}(\mathbf{c} - \boldsymbol{\mu}_m), \tag{2.8}$$

$$\boldsymbol{\Sigma}_B^* = \boldsymbol{\Sigma}_{B(-m,-m)} - \boldsymbol{\Sigma}_{B(-m,m)}\boldsymbol{\Sigma}_{B(m,m)}^{-1}\boldsymbol{\Sigma}_{B(m,-m)}. \tag{2.9}$$

To simulate $\mathbf{x}_t$ and $\boldsymbol{\theta}_t$ we note, from Equations (3.1), (3.2), and (3.3), that $\mathbf{x}_t = \mathbf{F}\boldsymbol{\theta}_t$ and

$$\begin{aligned} \mathbf{y}_t &= \mathbf{B}\mathbf{F}\boldsymbol{\theta}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim N(0, \mathbf{V}), \\ \boldsymbol{\theta}_t &= \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(0, \mathbf{W}). \end{aligned}$$

Thus, conditional on $\mathbf{B}$, $\mathbf{W}$, and $\mathbf{V}$, the model for $\mathbf{y}_t$ is a Gaussian dynamic linear model with $\boldsymbol{\theta}_t$ as the latent process [47, 58]. Therefore, we can use the well-known forward filter backward sampler (FFBS) algorithm to efficiently sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from their joint full conditional distribution [10, 19]. We can then use the fact that $\mathbf{x}_t = \mathbf{F}\boldsymbol{\theta}_t$ to obtain a sample of $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

In summary, the Gibbs sampler we propose proceeds in the following manner:

1. Set initial values for $\mathbf{X}, \boldsymbol{\theta}, \sigma_1^2, \ldots, \sigma_r^2, \mathbf{B}, \mathbf{W}$, and $\tau_1, \ldots, \tau_k$.

2. Simulate each observation variance σ_j^2, $j = 1, \ldots, r$, from its full conditional distribution given in Equation (2.5).

3. Simulate $\mathbf{W}$ from its full conditional distribution given in Equation (2.6).

4. Simulate each precision parameter τ_j, $j = 1, \ldots, k$, from its full conditional distribution given in Equation (2.7).

5. Simulate the free elements of matrix $\mathbf{B}$ from their joint full conditional distribution that is multivariate Gaussian with mean vector given in Equation (2.8) and covariance matrix given in Equation (2.9).

6. Simulate $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ using the FFBS algorithm.

7. Compute $\mathbf{x}_t = \mathbf{F}\boldsymbol{\theta}_t$, $t = 1, \ldots, n$.

8. Repeat steps from **2** to **7** until we have a posterior sample that is large enough after the Markov Chain has converged.

2.4.2 Model selection

An important component of Bayesian model selection is the predictive density, also known as integrated likelihood or marginal data density. For a k-factor DIFM, the predictive density is given by

$$\int\int\int\int\int p(\mathbf{y}|k, \mathbf{X}, \mathbf{B}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2) p(\mathbf{X}, \mathbf{B}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2|k) d\mathbf{X} d\mathbf{B} d\mathbf{W} d\boldsymbol{\tau} d\boldsymbol{\sigma}^2, \tag{2.10}$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_r^2)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$. In an extensive simulation study to compare several computation methods to evaluate the predictive density for factor models, Lopes and West [36] concluded that the Laplace-Metropolis estimator [34] provides a good approximation to the predictive density. Thus, here we perform model selection using predictive densities computed with the Laplace-Metropolis estimator.

As we explain below, the integral with respect to the latent factors $\mathbf{X}$ can be computed exactly with the Kalman filter. Let $p(\mathbf{y}|k, \mathbf{B}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2)$ be the integrated likelihood after integrating out $\mathbf{X}$. In addition, let d_k be the number of unknown parameters in $\mathbf{B}$, $\mathbf{W}$, $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}^2$. Then, the Laplace-Metropolis estimator of the predictive density for the k-factor DIFM is

$$(2\pi)^{\frac{d_k}{2}} |\boldsymbol{\Psi}|^{\frac{1}{2}} p(\mathbf{y}|k, \widehat{\mathbf{B}}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2) p(\widehat{\mathbf{B}}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2|k), \tag{2.11}$$

where $\boldsymbol{\Psi}$ is the posterior covariance matrix of a d_k-dimensional vector that contains all the

unknown parameters in $\mathbf{B}$, $\mathbf{W}$, $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}^2$ computed from the MCMC output. In addition, $\widehat{\mathbf{B}}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\tau}}$ and $\widehat{\boldsymbol{\sigma}}^2$ are the posterior modes of $\mathbf{B}, \mathbf{W}, \boldsymbol{\tau}$ and $\boldsymbol{\sigma}^2$ computed from the MCMC output.

To compute $p(\mathbf{y}|k, \mathbf{B}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2)$ using the Kalman filter, we note that the model for the latent factors $\mathbf{x}_t = \mathbf{F}\boldsymbol{\theta}_t$ conditional on $k, \mathbf{B}, \mathbf{W}, \boldsymbol{\tau}$, and $\boldsymbol{\sigma}^2$ can be written as a dynamic linear model [47]. Specifically, applying Equation (3.2) to Equation (3.1), we obtain

$$\mathbf{Y}_t^T = \mathbf{B}\mathbf{F}\boldsymbol{\theta}_t^T + \mathbf{V}.$$

By rewriting $\mathbf{BF}$ as $\mathbf{F}^{*T}$, we get

$$\mathbf{Y}_t^T = \mathbf{F}^{*T}\boldsymbol{\theta}_t^T + \mathbf{V}. \tag{2.12}$$

In addition, from Equation (3.3), $\boldsymbol{\theta}_t$ evolves as $\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$. Thus, Equations (3.3) and (2.12) imply a DLM with $\boldsymbol{\theta}_t$ as the latent state.

To apply the Kalman filter, let D_{t-1} be the information up to time *t - 1*. Then, from the Kalman filter, the posterior distribution at time $t-1$ is $\boldsymbol{\theta}_{t-1}|D_{t-1} \sim \mathrm{N}(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$. Combining this with Equation (3.3), the prior at time $t-1$ for $\boldsymbol{\theta}_t$ is $\boldsymbol{\theta}_t|D_{t-1} \sim \mathrm{N}(\mathbf{a}_t, \mathbf{R}_t)$ where $\mathbf{a}_t = \mathbf{G}\mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{G}\mathbf{C}_{t-1}\mathbf{G}^T + \mathbf{W}_t$. Thus, the predictive distribution at time $t-1$ for $\mathbf{Y}_t$ is $\mathbf{Y}_t|D_{t-1} \sim \mathrm{N}(\mathbf{f}_t, \mathbf{Q}_t)$ where $\mathbf{f}_t = \mathbf{F}^{*T}\mathbf{a}_t$ and $\mathbf{Q}_t = \mathbf{F}^{*T}\mathbf{C}_{t-1}\mathbf{F}^* + \mathbf{V}$. Finally, after $\mathbf{Y}_t$ is observed, the posterior distribution at time t of $\boldsymbol{\theta}_t$ can be obtained using Bayes Theorem yielding $\boldsymbol{\theta}_t|D_t \sim \mathrm{N}(\mathbf{m}_t, \mathbf{C}_t)$ where $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}^*\mathbf{Q}_t^{-1}$, $\mathbf{e}_t = \mathbf{Y}_{t,} - \mathbf{f}_t$, $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t\mathbf{e}_t$, and $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{Q}_t\mathbf{A}_t^T$. Thus, the integrated likelihood function with the latent common factors $\mathbf{X}$ integrated out can be computed as a product of a sequence of predictive densities

$$p(\mathbf{y}|k, \mathbf{B}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2) = \prod_{t=1}^{T} \mathrm{p}(\mathbf{Y}_t|D_{t-1}) = \prod_{t=1}^{T} \mathrm{N}(\mathbf{Y}_t|\mathbf{f}_t, \mathbf{Q}_t).$$

Therefore, the logarithm of this integrated likelihood function is

$$\log p(\mathbf{y}|k, \mathbf{B}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2) = -\frac{rn}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\log(|\mathbf{Q}_t|) - \mathbf{e}_t^T\mathbf{Q}_t^{-1}\mathbf{e}_t\right).$$

Let $\widehat{\boldsymbol{\Theta}} = (\widehat{\mathbf{B}}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2)$ be the posterior mode of $(\mathbf{B}, \mathbf{W}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2)$ computed from the MCMC output. Then, the Laplace-Metropolis approximation of the predictive density given in Equation (2.11) can be computed as

$$(2\pi)^{\frac{d_k}{2}} |\boldsymbol{\Psi}|^{\frac{1}{2}} p(\mathbf{y}|k, \widehat{\mathbf{B}}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2) \prod_{j=1}^{k} \exp\left(-\frac{1}{2\widehat{\tau}_j} (\widehat{\mathbf{B}}^*_{\cdot j} - \mathbf{h}_j)^T \mathbf{H}^*_j (\widehat{\mathbf{B}}^*_{\cdot j} - \mathbf{h}_j) \right)$$
$$\times \prod_{j=1}^{k} \frac{n_\tau s_\tau^2/2}{\Gamma(n_\tau/2)} \widehat{\tau}_j^{-(n_\tau+1)/2} \exp\left(-\frac{n_\tau s_\tau^2}{2\widehat{\tau}_j} \right) \prod_{i=1}^{r-1} \frac{n_\sigma s_\sigma^2/2}{\Gamma(n_\sigma/2)} \widehat{\sigma}_i^{-(n_\sigma+1)} \exp\left(-\frac{n_\sigma s_\sigma^2}{2\widehat{\sigma}_i^2} \right).$$

In what follows, we call this model selection criterion the Laplace-Metropolis predictive density.

2.5 Applications

In this section, we illustrate the application of our DIFM framework with analyses of two datasets. The first application considers a simulated dataset that allows us to verify that the DIFM is identifiable and our inference methods work properly. The second application considers monthly data from the PDODC dataset on the number of deaths caused by drug overdose from January 2015 to February 2021 by state in the contiguous United States.

In both applications presented in this section, we have used the following specifications for the priors of the latent process and of the hyperparameters. For the latent process at time $t = 0$, we assign a vague Gaussian prior with $\mathbf{m}_0 = \mathbf{0}_{2k}$ and $\mathbf{C}_0 = 10^4 \mathbf{I}_{2k}$. For the covariance matrix of the evolution equation, we assign a weakly informative inverse Wishart prior with $\mathbf{S}_W = 0.01\mathbf{I}$ and $n_W = 2k+2$. We note that $n_W = 2k+2$ is the smallest integer for which this inverse Wishart prior has a finite mean. In that case, the prior mean of $\mathbf{W}$ is equal to $\mathbf{S}_W$. For the idiosyncratic variances, we follow the suggestion of [36] and use a weakly informative prior with $n_\sigma = 2.2$ and $n_\sigma s_\sigma^2 = 0.1$. Similarly, for the spatial correlation parameter τ, we assign a weakly informative prior with $n_\tau = 2.2$ and $n_\tau s_\tau^2 = 0.1$.

2.5.1 Simulated Dataset

To assess the identifiability of our DIFM model and the correctness of our proposed estimation approach, we analyze a simulated dataset that mimics the real dataset considered in Section 2.5.2. Specifically, we simulate a dataset using a 3-factor DIFM with true values of the parameters equal to the estimated values from the real dataset.

We have run the MCMC algorithm proposed in Section 2.4.1 for 50,000 iterations and discarded the first 5,000 iterations as burn-in. In addition, we have performed model selection as described in Section 2.4.2 to select the number of factors.

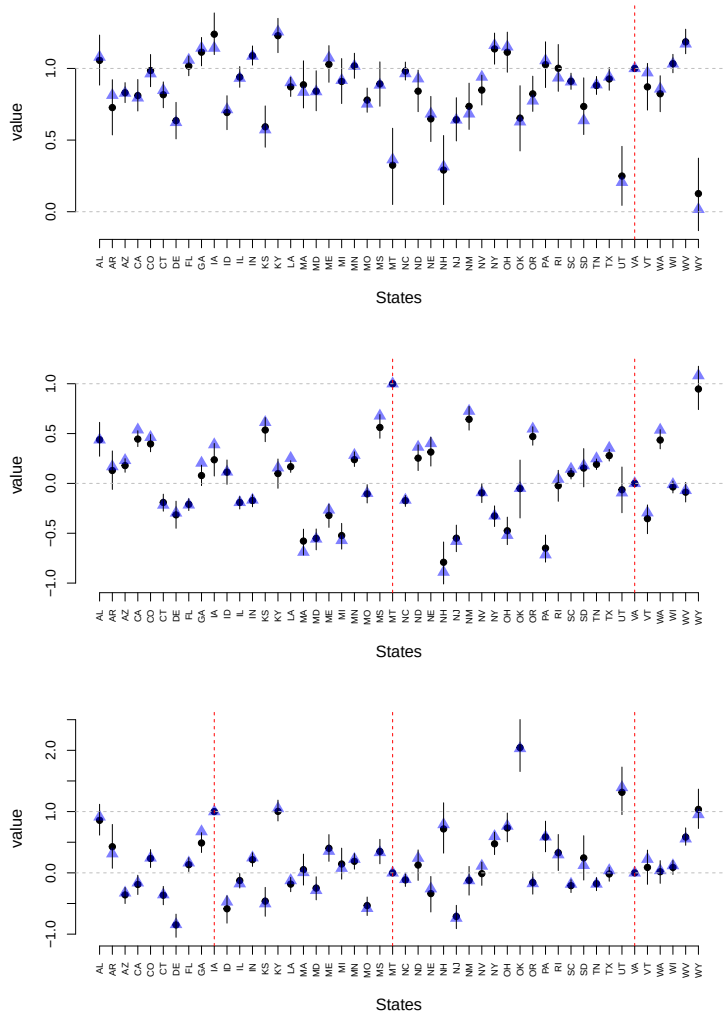


Figure 2.3: Simulated dataset – factor loadings for the 3-factor DIFM: true value (blue triangle), posterior mean (black circle), and 95% credible interval (black vertical line). Vertical dashed red lines indicate loadings fixed by the hierarchical structural constraint.

Figure 2.3 displays the true value (blue triangle), posterior mean (black circle), and 95% credible interval (black vertical line) for the factor loadings. In addition, vertical dashed red lines indicate loadings fixed by the hierarchical structural constraint. To set the constraint, the first three states were ordered as Virginia, Montana, and Iowa. Importantly, Figure 2.3 shows that our proposed point and interval estimation methods work well.

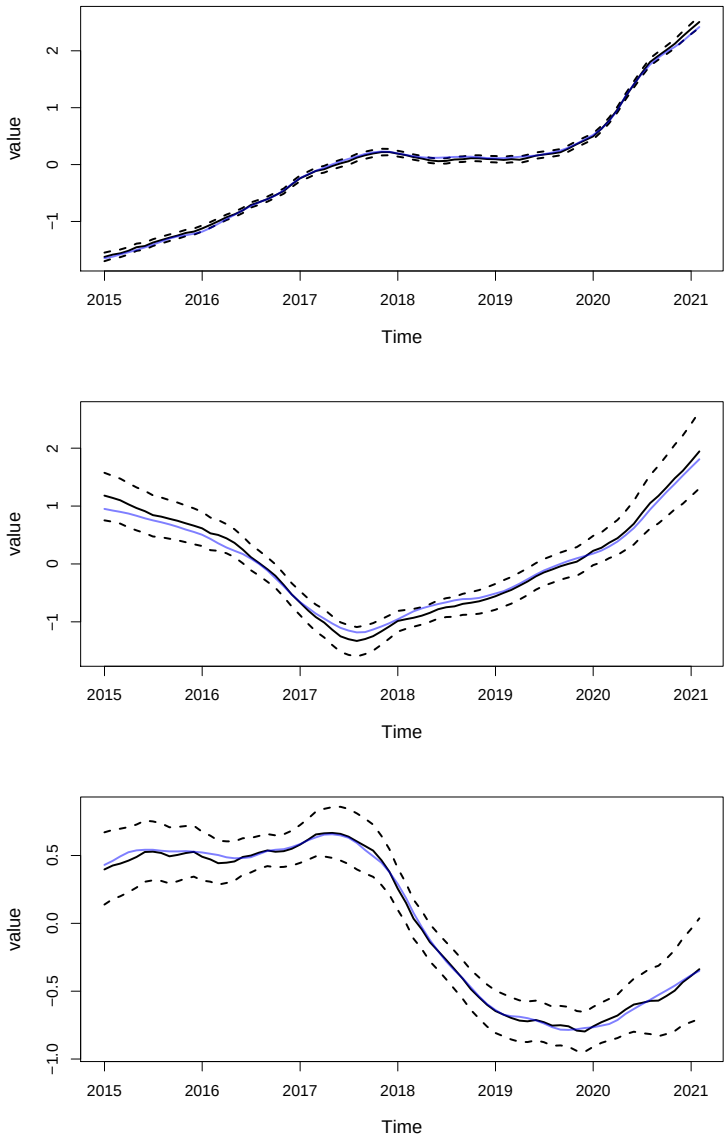


Figure 2.4: Simulated dataset – common factors for the 3-factor DIFM: true value (blue line), posterior mean (black solid line), and 95% credible interval (black dashed lines).

Figure 2.4 shows the true value (blue line), posterior mean (black solid line), and 95% credible interval (black dashed lines) for the three common factors. The uncertainty in the estimation of the first common factor is much lower than the uncertainty in the estimation of the other two common factors. In addition, for all three common factors, the posterior mean is close to the true value. Finally, the credible intervals contain the true values of the common factors most of the time. Therefore, our estimation procedure is also working well for the common factors.

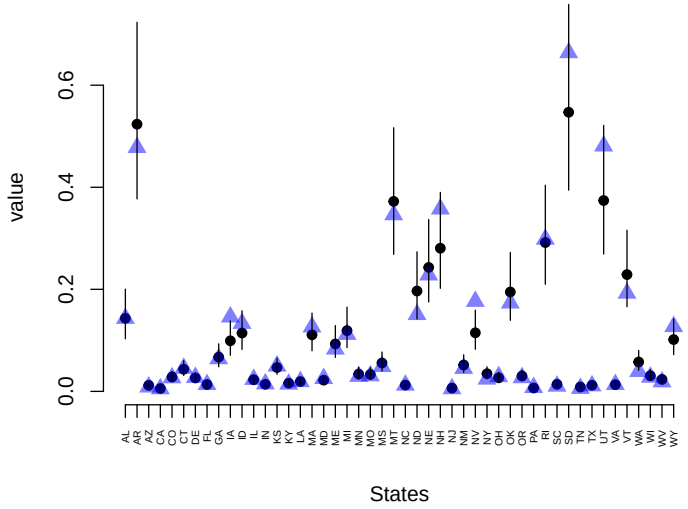


Figure 2.5: Simulated dataset – observational error variance for the 3-factor DIFM: true value (blue triangle), posterior mean (black circle), and 95% credible interval (black vertical line).

Figure 2.5 presents the true value (blue triangle), posterior mean (black circle), and 95% credible interval (black vertical line) for the idiosyncratic variances. Note that we have simulated this dataset with the true idiosyncratic variances assuming a wide variety of values. Specifically, most of the states have small true idiosyncratic variance under 0.2, while there are some few exceptions that have idiosyncratic variance larger than 0.4. In the estimation of these idiosyncratic variances, when the true variance is small the credible interval is narrow and the posterior mean is nearly identical to the true value. When the true idiosyncratic variance is large, there is more uncertainty in its estimation. Finally, Figure 2.5 shows that our inference approach provides appropriate uncertainty quantification.

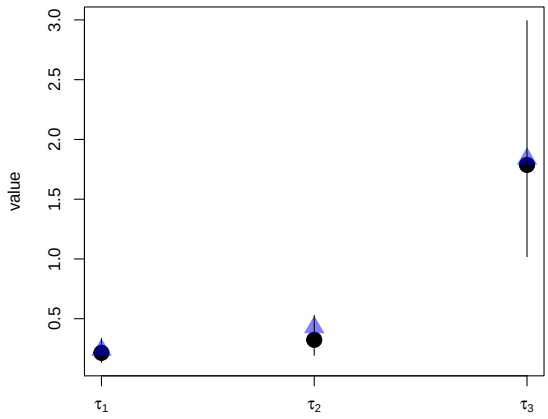


Figure 2.6: Simulated dataset – spatial dependence parameter τ_j, $j = 1, 2,$ and 3, for the 3-factor DIFM: posterior mean (black circle), 95% credible interval (black line) and true value (blue triangle).

Figure 2.6 shows posterior mean (black circle), 95% credible interval (black line) and true value (blue triangle) for τ_j, $j = 1, 2, 3$, the spatial dependence parameter for the factor loadings of the jth factor. The posterior means are close to the true values, showing that our estimation procedure works well. In addition, smaller true values of τ_j have narrower credible intervals. Further, we note that larger values of τ_j imply stronger spatial dependence. Thus, our estimation procedure is able to capture the fact that for this simulated dataset the factor loadings of the third factor have stronger spatial dependence than the factor loadings of the first and second factors.

Table 2.1: Simulated dataset – Logarithm of Laplace-Metropolis predictive density (log PD) for DIFMs with number of factors from 1 to 6.

# factors	1	2	3	4	5	6
log PD	-6509.2	-1155.5	**-303.1**	-470.6	-631.3	-699.1

To perform model selection, we have computed the predictive density using the Laplace-Metropolis approximation as we have described in Section 2.4.2. Table 2.1 presents the logarithm of the Laplace-Metropolis predictive density of DIFMs with 1, 2, 3, 4, 5, and 6 factors. The 3-factor model, which was the true model used to simulate this dataset, has the largest predictive density. Therefore, for this simulated dataset, our model selection procedure chooses the true model with the correct number of factors.

2.5.2 Real Dataset

We illustrate the use of our framework with an application to the PDODC dataset from the CDC. Specifically, we analyze the monthly number of deaths per 100,000 people by drug overdose in each of the 48 contiguous states of the United States from January 2015 to February 2021. To normalize and variance-stabilize the data, we consider the square root of the monthly standardized number of deaths per 100,000 inhabitants. All results presented in this section are based on 100,000 iterations of the MCMC algorithm presented in Section 2.4.1 discarding 10,000 iterations as burn-in.

Table 2.2: Real data – Logarithm of Laplace-Metropolis predictive density (log PD) for DIFMs with number of factors from 1 to 10.

# factors	1	2	3	4	5	6	7	8	9	10
log PD	-6470.1	-1011.4	-230.1	69.2	422.4	615.3	798.3	871.4	**945.2**	929.4

We have fitted to this dataset DIFMs with 1 to 10 factors. The order of the variables for the hierarchical structural constraint has been decided similarly to the exploratory data analysis presented in Section 2.2. Specifically, the order is Virginia, Montana, Iowa, Alabama, Utah, Nebraska, New Hampshire, North Dakota, Arkansas, and Ohio. To choose the number of factors, we performed model selection as described in Section 2.4.2. Table 2.2 presents the Laplace-Metropolis approximation of the predictive density for the several competing DIFMs. The best model according to the Laplace-Metropolis predictive density is the DIFM with 9 factors. Therefore, henceforth we present results for the DIFM with 9 factors.

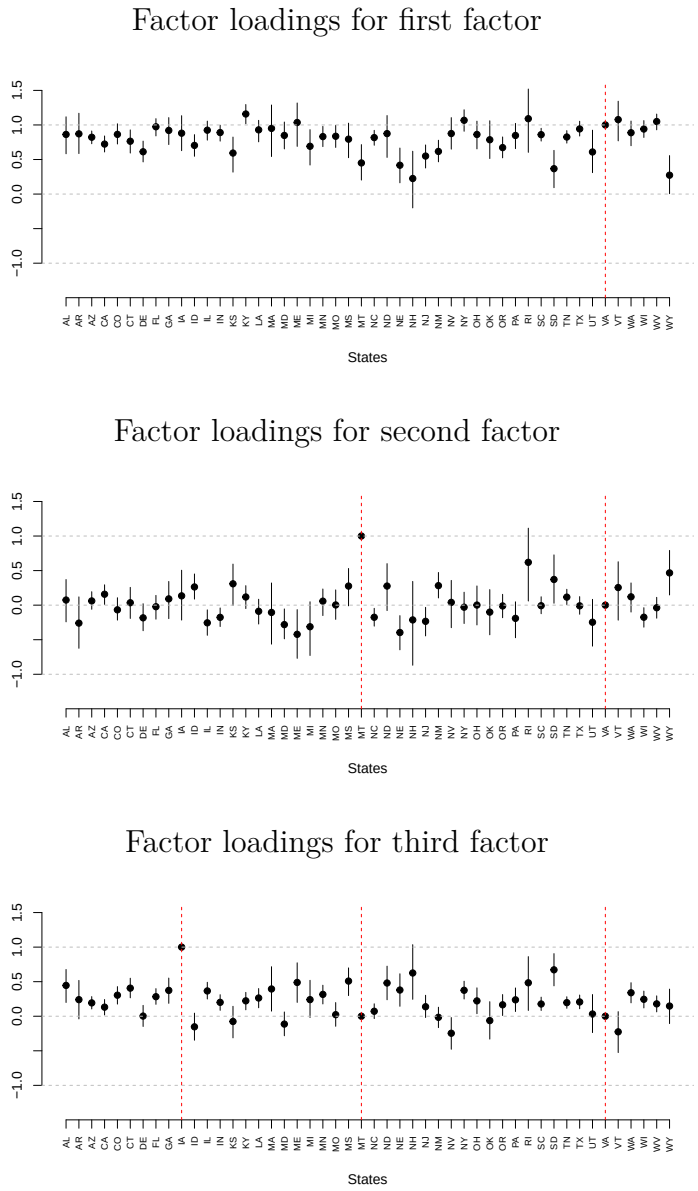


Figure 2.7: Real dataset – factor loadings for the 9-factor DIFM: posterior mean (black circle), and 95% credible interval (black vertical line). Vertical dashed red lines indicate loadings fixed by the hierarchical structural constraint.

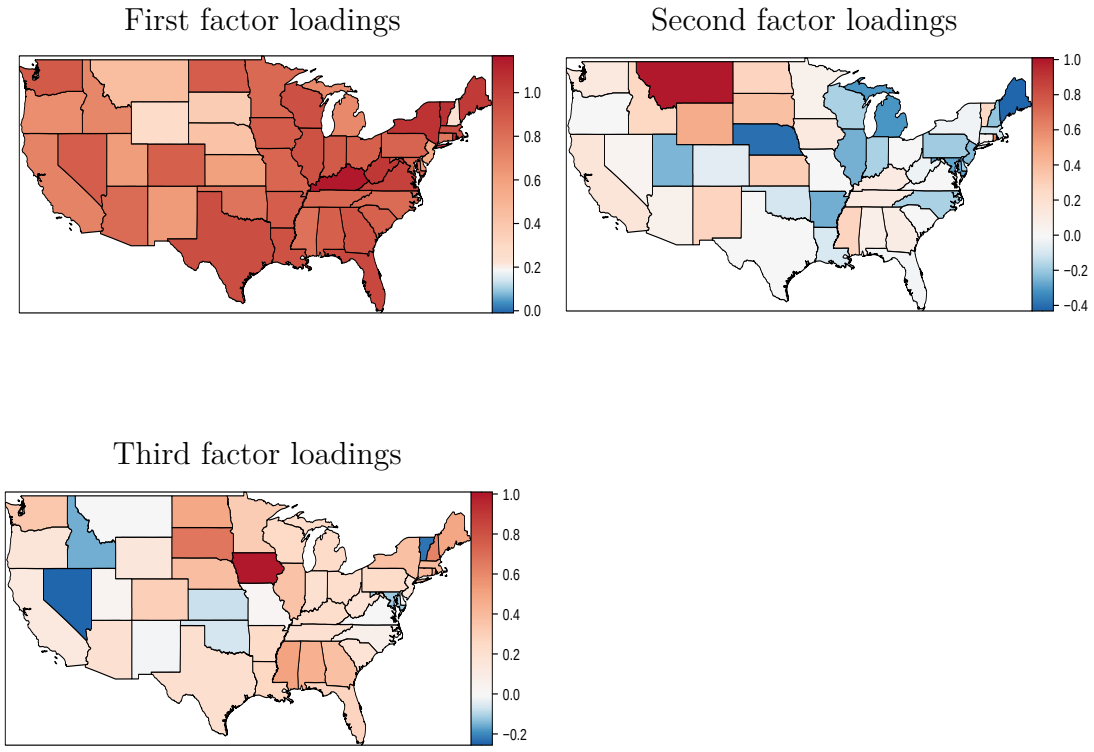


Figure 2.8: Real dataset – maps of factor loadings for the first three factors of 9-factor DIFM.

Figure 2.7 displays posterior mean (black circle), and 95% credible intervals (black vertical line) of the factor loadings of the first 3 factors of the 9-factor DIFM. Vertical dashed red lines indicate loadings fixed by the hierarchical structural constraint. We note that these same factor loadings can be put on maps as shown in Figure 2.8. While Figure 2.7 provides information about the level and uncertainty related to each factor loading, Figure 2.8 provides information about the spatial distribution of the factor loadings. In particular, the factor loadings for the first factor have strong spatial dependence. For example, there is a cluster of larger factor loadings in the northeast, and there is another cluster in the west formed by Montana, Wyoming, South Dakota, and Nebraska. The factor loadings for the second factor have some spatial dependence. For example, there is a cluster of negative factor loadings formed by Michigan, Wisconsin, Illinois and Indiana. The factor loadings for the third factor also have spatial dependence. For example, there is a cluster of positive factor loadings in the Southeast.

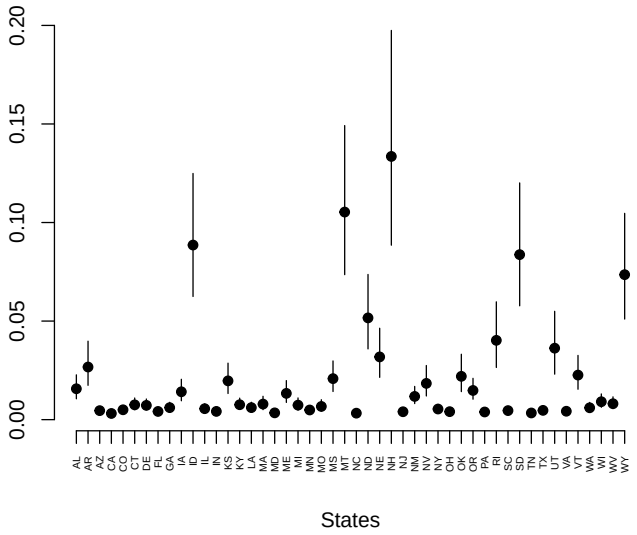


Figure 2.9: Real dataset – idiosyncratic variance for each state based on the 9-factor DIFM: posterior means (black circle) and 95% credible intervals (black vertical line).

Figure 2.9 shows the posterior means (black circle) and 95% credible intervals (black vertical line) of the idiosyncratic variances for each state based on the 9-factor DIFM. Most of the states have idiosyncratic variance smaller than 0.1. In addition, the states with large idiosyncratic variances have smaller population sizes, which is intuitive because raw standardized mortality ratios for regions with smaller populations usually have larger variances.

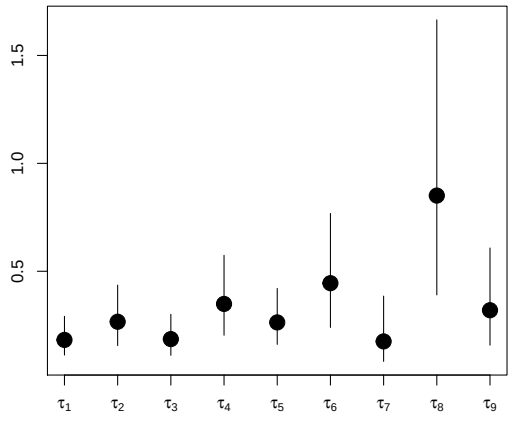


Figure 2.10: Real dataset – spatial dependence parameters for the 9-factor DIFM: posterior means (black circle) and 95% credible intervals (black line).

Figure 2.10 shows posterior means (black circle) and 95% credible intervals (black line) of

the spatial correlation parameters τ_j of the 9-factor DIFM. The parameter τ_j controls the strength of the spatial correlation within the jth factor, with larger values implying stronger spatial correlation. Thus, in this application the factor loadings of the eighth factor have the strongest spatial correlation.

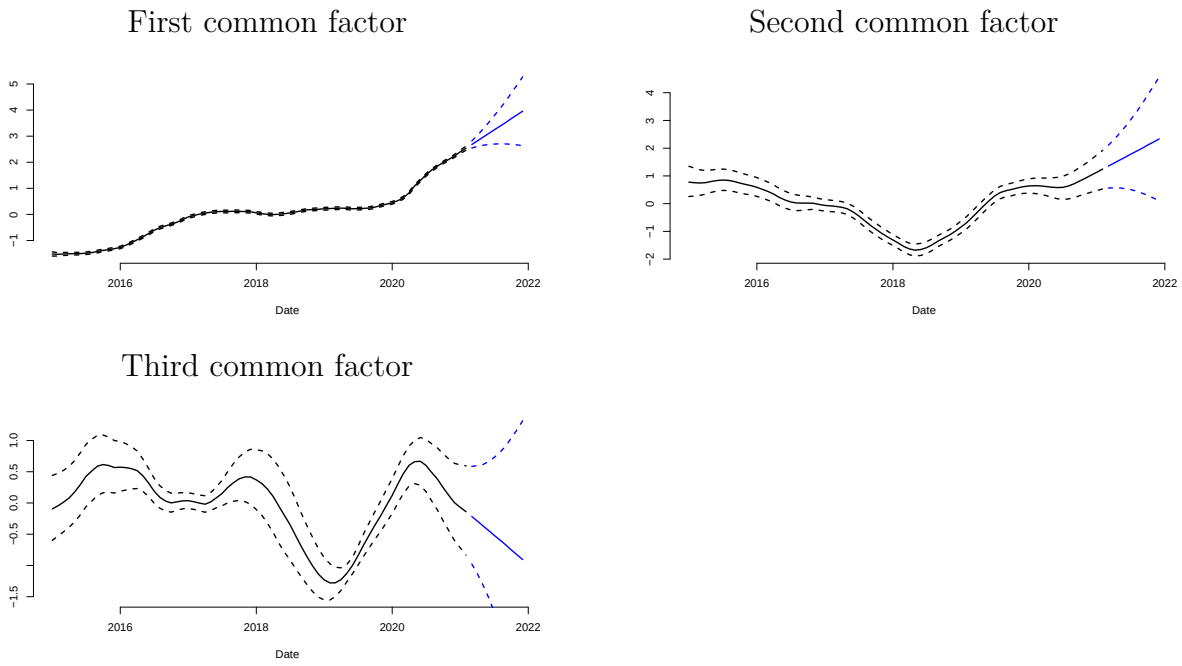


Figure 2.11: Real dataset 9-factor DIFM – first three common factors and 10-step ahead forecasts: posterior means (black solid line), 95% credible intervals (black dashed lines), 10-step ahead predictive means (blue solid line), and 10-step ahead 95% predictive intervals (blue dashed lines).

Figure 2.11 shows the posterior means (black solid line), 95% credible intervals (black dashed lines), 10-step ahead predictive means (blue solid line), and 10-step ahead 95% predictive intervals (blue dashed lines) of the first three common factors of the 9-factor DIFM. The black solid line starts from January 2015 and ends in February 2021. The forecasts are from March 2021 to December 2021. The first common factor was steadily increasing until about the end of 2017, when it stabilized until the end of 2019. In the beginning of 2020, the first common factor started increasing again due to the COVID pandemic. The second common factor was decreasing until the beginning of 2018, when it started increasing. In the middle of 2019, the second common factor stabilized. After the start of the COVID pandemic, the second common factor started increasing again. The third common factor seems to exhibit

a quasi-cyclical behavior that warrants further investigation that is outside the scope of this manuscript.

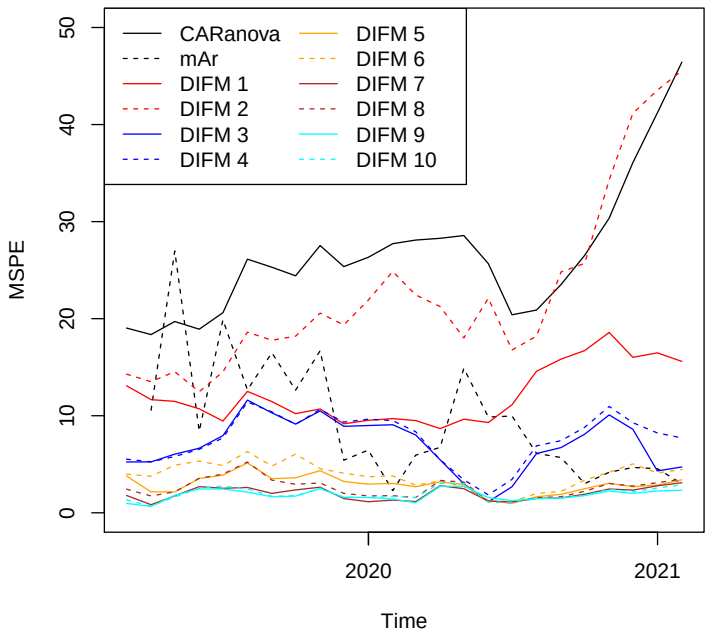


Figure 2.12: MSPE – Mean squared prediction error of one-step-ahead predictions from the CAR ANOVA model (black solid line), multivariate autoregressive(1) (black dashed line), and DIFMs with number of factors varying from 1 to 10. The forecasted timepoints are from March 2019 to February 2021.

To further evaluate DIFMs, we compare their predictive performance against that of the CAR ANOVA model [30] implemented with the `CARBayesST` R package [32]. Specifically, Figure 2.12 shows the mean squared prediction error (MSPE) of one-step-ahead predictions from the CAR ANOVA model and DIFMs with number of factors varying from 1 to 10 for the period from March 2019 to February 2021. The black solid line shows the MSPE through time of the CAR ANOVA, while the colored lines show the MSPEs of the several DIFMs. For DIFMs with 2 or more factors, the MSPE is always smaller than the MSPE of the CAR ANOVA model. The MSPE of the DIFMs usually drops as we add more factors up to 8 factors. However, the MSPEs of the DIFMs with 8, 9, and 10 factors are about the same. This confirms and justifies the choice of the 9-factor DIFM by the Laplace-Metropolis predictive density. In particular, we note that the overall MSPE of the 9-factor DIFM is

2.07 whereas the MSPE of the CAR ANOVA model is 26.47. Therefore, when compared to the CAR ANOVA model, our chosen DIFM provides much better predictions.

2.6 Conclusions

We have introduced the DIFM, the Dynamic ICAR Spatiotemporal Factor Model that is useful for the analysis of spatiotemporal areal data. Our DIFM assumes that the vector of observations that collects the spatial areal data at each time point can be written as a matrix of factor loadings times a vector of common factors plus a vector of errors. Each row of the matrix of factor loadings corresponds to a subregion. Thus, each column of the matrix of factor loadings corresponds to a vectorized map of the region of interest. Hence, to account for spatial correlation amongst the subregions, our DIFM assumes for each column of the matrix of factor loadings an intrinsic conditional autoregressive model. In addition, the vector of common factors evolves through time according to a dynamic linear model. Typically, the vector of common factors has a much lower dimension than the number of subregions. As a consequence, our DIFM may achieve substantial dimension reduction. Finally, we have developed an efficient Gibbs sampler with an embedded forward filter backward sampler for posterior exploration.

We have presented an application of DIFMs to the number of deaths caused by drug overdose by state in the contiguous United States. In this application, the best DIFM for 48 states has nine factors. The first factor represents an overall mean. The first factor had an increasing trend up to the end of 2017, when it stopped increasing and stabilized, and then it started increasing again at the beginning of 2020 due to the COVID pandemic. The other eight factors represent various spatial clusters. As a salient feature, our DIFM framework allows temporal forecasting of each common factor and of the original spatiotemporal process.

We have compared the predictive performance of DIFMs with varying number of factors to the predictive performance of the CAR ANOVA model [30] implemented with the `CARBayesST` R package [32]. We have found that DIFMs with 2 or more factors have smaller mean squared prediction error than the CAR ANOVA model. We note that the Laplace-Metropolis predictive density that we propose chooses, among DIFMs, the 9-factor DIFM.

Importantly, when compared to the CAR ANOVA model, the 9-factor DIFM provides a 92% reduction in the mean squared prediction error.

An important future research direction would be to extend our DIFMs to the case of non-Gaussian observations. Here, we have presented a case study with count data where the number of counts was large enough for the Gaussian distribution to be a good approximation to model the square root of the number of counts. However, such an approximation would not be reasonable for cases of low counts. Therefore, a research direction of great practical importance would be to extend our framework to observations with distributions in the exponential family. Such an extension would allow applications of DIFMs to spatiotemporal analyses of diseases with low counts.

Chapter 3

Bayesian Clustering Factor Models

3.1 Introduction

Factor models are widely used to identify meaningful latent structures in multivariate data. On the other hand, Gaussian mixture models can identify latent clusters. In practice, researchers often find clusters in datasets with high dimensional multivariate data by first performing Principal Component Analysis (PCA) or factor analysis (FA) to reduce the dimension of the problem, followed by clustering the data using a k-means method [14]. Here, we propose a more formal approach combining Bayesian factor models with Gaussian mixture models to concomitantly reduce dimension and find clusters. We call our models Bayesian Clustering Factor Models (BCFM).

We assume there are latent factors derived through linear relationships of the covariates. In general, the number of factors is smaller than the dimension of the data. Therefore, factor models can significantly reduce the dimensionality. In addition, Bayesian inference can add great flexibility to the factor models. Bayesian factors model can explain multivariate dependence and hierarchical structures of variables. For example, it can explain spatial panel data with temporal components [33] and be applied to categorical data [9]. In our model, the structures of mean and covariance of the common factors are different by clusters, but the factor loadings remain identical. To ensure the clusters do not swap during the MCMC process, we set the covariance structure of the largest cluster in the initial step as a diagonal matrix. The other clusters are allowed to have non-diagonal covariance structures. We define the common factors with respect to the largest cluster in this manner. The clusters do not interchange and have the same interpretations throughout the Gibbs sampler. When the clusters have different factor loading matrices, the interpretations of the factors would be different across the groups. In the applications we consider, it is more reasonable to assume

that the interpretations of the factors are identical across the clusters. Therefore, we set the clusters to share the same factor loadings.

Clustering is one of the most common unsupervised methods. We assign the classes by the distance among the samples. The observations of the same cluster are adjacent, while those of different clusters are distant. An example of clustering methods is k-means clustering, which iteratively defines the cluster centroids until the sum of distances among the observations within the group is minimized [4]. Another popular method is hierarchical clustering, which determines the link of the observations in a hierarchical order by the definition of distances [41]. Traditional k-means and hierarchical clustering are applicable when the data consists of quantitative information. For categorical information, Latent Cluster Analysis (LCA) is an effective way to group the subjects [55].

Finding the appropriate number of clusters and factors in Bayesian models is challenging. One application is to use infinite factor models, which employs multiplicative gamma priors [7, 40] or fit a model with overfitting the data with a large number of clusters [44]. We normally set Gaussian model to common factors, but this can be replaced with t-distribution to account for large variance [3]. Bayesian factor models may have identifiability issues on factor loadings and common factors. In addition, many mixture models have label switching issues [53]. While these are not problematic in predictions, they affect the interpretation of the factors. We can solve the issue by putting constraints on the model structure [54] or by setting restrictions during the computational steps [45].

Our BCFM requires to prespecify the number of factors and clusters. To achieve this, k-means clustering and silhouette score are helpful to explore the number of clusters to begin BCFM. Principal Component Analysis (PCA) is a good method to preview the factors of the dataset. We run PCA before BCFM to find the appropriate number of factors and learn the amount of variance factors can explain [24, 26]. We run BCFM with different settings and compare how the subjects are allocated to each cluster. To examine the most appropriate model, we develop the Laplace-Metropolis approximation of the marginal density and the BIC with integrated likelihood.

We present the work of BCFM through simulated datasets, the Opioid Use Disorder (OUD) dataset, and the breast cancer molecular subtype gene expression dataset. We show a single simulated dataset to demonstrate the posterior distribution of the MCMC sample. In ad-

dition, we simulate 300 datasets to illustrate the results of 3 different settings and compare our model with the Bayesian mixture model proposed in [44]. In section **??**, BCFM with BIC with integrated likelihood is better at selecting the correct models. In the OUD data, we explore the data to find how the subgroups of subjects are defined. In the breast cancer dataset, we discover how the two assessment criteria choose the number of clusters and compare the cluster assignments with true subtypes.

The following sections are organized as follows. Section 3.2 introduces the formula and the structure of the model parameters we propose. Section 3.3 explains further how the priors and posteriors illustrated in Section 3.2 are applied to the model. The results of the simulated data and the real data are presented in Section 3.5. Finally, we discuss the conclusions of the paper in Section 3.6.

3.2 Model Specification

3.2.1 Bayesian Clustering Factor Model

We consider a setting with multivariate R variables observed for each of n subjects. In addition, we assume that we can explain the dependence structure among the R variables with a much smaller number F of latent variables or factors. Further, we assume that in this smaller F-dimensional latent space, the subjects may be clustered into K clusters. This section describes the model we propose for performing this concomitant dimension reduction and clustering.

Let $\mathbf{y}_i$ be an R-dimensional vector with the R observed variables from subject i. We assume the following factor model with F factors

$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i + \mathbf{v}_i, \tag{3.1}$$

where $\mathbf{B}$ is an $R \times F$ matrix of factor loadings, and $\mathbf{x}_i$ is an F-dimensional vector of common factor for subject i. The error vector is $\mathbf{v}_i \sim \mathbf{N}(\mathbf{0}, \mathbf{V})$ with $\mathbf{V} = diag(\sigma_1^2, \ldots, \sigma_R^2)$ and $\sigma_1^2, \ldots, \sigma_R^2$ are the idiosyncratic variances. Note that Equation (3.1) encodes a dimension reduction from dimension R to dimension F. To ensure the identifiability of the model,

we assume that the matrix of factor loadings $\mathbf{B}$ follows a hierarchical structural constraint [23, 47**?**]. Specifically, $\mathbf{B}$ has the form

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ b_{2,1} & 1 & 0 & \dots & 0 \\ b_{3,1} & b_{3,2} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{F,1} & b_{F,2} & b_{F,3} & \dots & 1 \\ b_{F+1,1} & b_{F+1,2} & b_{F+1,3} & \dots & b_{F+1,F} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{R,1} & b_{R,2} & b_{R,3} & \dots & b_{R,F} \end{bmatrix}.$$

Thus, the matrix of factor loadings $\mathbf{B}$ is a lower triangular matrix with main diagonal elements all equal to 1. Each row of $\mathbf{B}$ corresponds to an observed variable, and each column corresponds to a common factor. The order of the variables should be chosen carefully because of the properties of the hierarchical structural constraints.

We assume that the common factors $\mathbf{x}_i$ follow a Gaussian mixture model. This allows the ith subject to be allocated to one of the K clusters. Let z_i indicate the cluster subject i belongs to. Then, given $z_i = k$, the common factor x_i has the Gaussian conditional distribution

$$\mathbf{x}_i | z_i = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k), \tag{3.2}$$

where $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kF})'$ is the mean vector and $\boldsymbol{\Omega}_k$ is the covariance matrix of the common factors of cluster k. Let $\boldsymbol{\mu}_{.j} = (\mu_{1j}, \dots \mu_{Kj})'$ be the vector that contains the jth element of the mean vectors across clusters. Let the probability of a randomly selected subject belonging to cluster k be $P(z_i = k) = p_k$. Then, the Gaussian mixture model for $\mathbf{x}_i$ is

$$\mathbf{x}_i \sim \sum_{k=1}^{K} p_k N(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k). \tag{3.3}$$

Note that to make the model identifiable, we impose a constraint that the first cluster has a diagonal covariance matrix.

3.2.2 Priors

We consider conditionally conjugate priors for the factor loadings. Specifically, the priors of the unconstrained elements in the lth factor are assumed to be independent and identically normally distributed with mean 0 and variance τ_l, $l = 1, \ldots, F$. Thus, the elements of the rth row of matrix $\mathbf{B}$, $r > F$, follow a Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathrm{T} = diag(\tau_1, \ldots, \tau_F)$. When $2 \leq r \leq F$, the unconstrained elements of the rth row of $\mathbf{B}$ follow *a priori* a Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $diag(\tau_1, \ldots, \tau_{r-1})$. We assume for the variance of the lth factor τ_l an inverse gamma prior $IG(n_\tau/2, n_\tau s_\tau^2/2)$. Similarly, we assume for the idiosyncratic variance σ_r^2 an inverse gamma prior $IG(n_\sigma/2, n_\sigma s_\sigma^2/2)$.

From Equation (3.3), the cluster assignment variable z_i is assigned a discrete distribution with $P(z_i = k) = p_k$, $k = 1, \ldots, K$. Thus, we assign for the probability vector $\mathbf{p} = (p_1, \ldots, p_K)$ a conditionally conjugate Dirichlet prior $Dirichlet(\alpha_1, \ldots, \alpha_K)$. In addition, we assign for the mean vector of the kth cluster, $k = 1, \ldots K$, the prior $\boldsymbol{\mu}_k \sim N(\mathbf{m}_k, \mathbf{C}_k)$, where $\mathbf{m}_k$ and $\mathbf{C}_k$ are known. Further, since the covariance matrix $\boldsymbol{\Omega}_1$ of the first cluster is assumed to be diagonal, we assign inverse gamma prior to the lth diagonal element, $\{\Omega_1\}_{ll} \sim IG(n_{\omega l}/2, n_{\omega l} s_{\omega l}^2/2)$. Finally, we assume conditionally conjugate inverse Wishart prior $\boldsymbol{\Omega}_k \sim IW(\nu + F, \boldsymbol{\Psi}_k)$ for covariance matrix $\boldsymbol{\Omega}_k$, $k = 2, \ldots, K$.

Prior Hyperparameters

To facilitate the assignment of priors, we first standardize the data. Thus, following the recommendations of [36] and [47], we choose $n_\sigma = 2.2$ and $n_\sigma s_\sigma^2 = 0.1$. This choice implies the idiosyncratic variances have a prior mean equal to 0.5 and infinite prior variance. Thus, this is a vague prior. In addition, because the data are standardized, and following the recommendations of [47], we choose $n_\tau = 1$ and $n_\tau s_\tau^2 = 1$.

Inspired by unit information priors [27] for model selection, we assign weakly informative priors for the cluster parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. Specifically, we determine the hyperparameters for these priors with a preliminary principal component analysis followed by k-means clustering. Let $\widehat{\mathbf{B}}$ be an $R \times F$ matrix where the columns are the first F eigenvectors from the PCA analysis. In addition, let $\widehat{\mathbf{x}}_i$ be the principal component of the ith observation vector $\mathbf{y}_i$.

We use the principal components $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n$ to determine the prior hyperparameters for $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. Note that $\widehat{\mathbf{B}}$ does not follow the hierarchical constraint. Let $\mathbf{M}$ be a matrix such that $\widehat{\mathbf{B}}^* = \widehat{\mathbf{B}}\mathbf{M}$ satisfies the hierarchical structural constraint. Let $\widehat{\mathbf{x}}_i^* = \mathbf{M}^{-1}\widehat{\mathbf{x}}_i$. Thus, $\widehat{\mathbf{B}}^*\widehat{\mathbf{x}}_i^* = \widehat{\mathbf{B}}\widehat{\mathbf{x}}_i$.

Next, we apply k-means clustering to $\widehat{\mathbf{x}}_1^*, \ldots, \widehat{\mathbf{x}}_n^*$. We order the clusters in decreasing order of cluster sizes. Let $\mathbf{S}_1, \ldots, \mathbf{S}_G$ be the sample covariance matrices of the transformed principal components $\widehat{\mathbf{x}}_i^*$ within each of the G clusters identified by the k-means clustering. Denote the LDL decomposition of $\mathbf{S}_1$ by $\mathbf{S}_1 = \mathbf{L}_1\mathbf{D}_1\mathbf{L}_1'$ where $\mathbf{L}_1$ is a lower triangular matrix with diagonal elements equal to 1 and $\mathbf{D}_1$ is a diagonal matrix. Let $\widetilde{\mathbf{x}}_i = \mathbf{L}_1^{-1}\widehat{\mathbf{x}}_i^*$. Then, the sample covariance matrix of the transformed principal components $\widetilde{\mathbf{x}}_i$ that belong to the first cluster is $\mathbf{L}_1^{-1}\mathbf{S}_1(\mathbf{L}_1^{-1})' = \mathbf{D}_1$, which is a diagonal matrix. In addition, note that $\widetilde{\mathbf{B}} = \widehat{\mathbf{B}}^*\mathbf{L}_1$ also satisfies the hierarchical structural constraint and $\widetilde{\mathbf{B}}\widetilde{\mathbf{x}_i} = \widehat{\mathbf{B}}\widehat{\mathbf{x}}_i$. Therefore, $\widetilde{\mathbf{B}}$ and $\mathbf{D}_1$ satisfy the constraints we use in our BCFM.

Let $\widetilde{Z}_k$ be the kth cluster identifed by the k-means clustering. Let n_k be the number of observations in cluster $\widetilde{Z}_k$. Thus, the hyperparameters in the prior for $\boldsymbol{\mu}_k$ are $\mathbf{m}_k = n_k^{-1}\sum_{i\in Z_k}\widetilde{\mathbf{x}}_i$ and $\mathbf{C}_k = \mathbf{L}_1^{-1}\mathbf{S}_k(\mathbf{L}_1^{-1})'$. In addition, the hyperparameters in the prior for the lth diagonal element of $\boldsymbol{\Omega}_1$ are $n_{\omega l} = 2$ and $n_{\omega l}s_{\omega l}^2 = \{\mathbf{D}_1\}_{ll}$. Further, the hyperparameters in the prior for $\boldsymbol{\Omega}_k$, $k = 2, \ldots, K$, are $\nu = 2$ and $\boldsymbol{\Psi}_k = \mathbf{L}_1^{-1}\mathbf{S}_k(\mathbf{L}_1^{-1})'$. This weakly informative prior solves issues of identifiability due to label switching [53].

The cluster mean of the kth cluster $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, is assigned a Gaussian prior $N(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*)$. Let $\widehat{\boldsymbol{\mu}}_k$ be the sample mean of the kth cluster obtained with PCA and k-means. Then, taking into account the LDL decomposition, the prior mean of $\boldsymbol{\mu}_k$ is $\boldsymbol{\mu}_k^* = \mathbf{L}_1^{-1}\mathbf{M}^{-1}\widehat{\boldsymbol{\mu}}_1$.

The choice of hyperparameters for the prior distribution of the probabilities $p_1, \ldots, p_K$, has to be done in a careful manner. In particular, while usual non-informative priors for probabilities assign $(\alpha_1, \ldots, \alpha_K) = (0.5, \ldots, 0.5)$ or $(\alpha_1, \ldots, \alpha_K) = (1, \ldots, 1)$, for mixture models these prior choices allow clusters with very low probability that may become empty during the MCMC algorithm. Thus, to keep the probabilities $p_1, \ldots, p_K$ away from zero, we assume prior $(\alpha_1, \ldots, \alpha_K) = (2, \ldots, 2)$.

3.3 Statistical Inference

3.3.1 Posterior Exploration

We propose a Gibbs sampler [22] to explore the posterior distribution of the BCFM parameters. This section presents the full conditional distributions of the parameters.

The full conditional of the common factor $\mathbf{x}_i$ depends on the cluster assignment z_i. Given $z_i = k$, the full conditional of $\mathbf{x}_i$ is $N(\mathbf{m}_i, \mathbf{A}_i), i = 1, \ldots, n$, where

$$\begin{aligned} \mathbf{A}_i = & \left(\boldsymbol{\Omega}_k^{-1} + \mathbf{B}'\mathbf{V}^{-1}\mathbf{B}\right)^{-1}, \\ \mathbf{m}_i = & \left(\boldsymbol{\Omega}_k^{-1} + \mathbf{B}'\mathbf{V}^{-1}\mathbf{B}\right)^{-1}\left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{y}_i + \boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k\right). \end{aligned} \tag{3.4}$$

The mean vector $\boldsymbol{\mu}_k$ of the kth cluster, $k = 1, \ldots, K$, has the following Gaussian full conditional distribution

$$\boldsymbol{\mu}_k|\mathbf{Y}, \mathbf{X} \sim N\left(\left(\boldsymbol{\Omega}_k^{*-1} + n_k\boldsymbol{\Omega}_k^{-1}\right)^{-1}\left(\boldsymbol{\Omega}_k^{*-1}\boldsymbol{\mu}_k^* + n_k\boldsymbol{\Omega}_k^{-1}\bar{\mathbf{X}}_k^*\right), \left(\boldsymbol{\Omega}_k^{*-1} + n_k\boldsymbol{\Omega}_k^{-1}\right)^{-1}\right), \tag{3.5}$$

where $\bar{\mathbf{X}}_k^*$ is the mean of common factors for the observations that belong to the kth cluster and n_k is the size of the kth cluster.

Recall that the first cluster covariance matrix $\boldsymbol{\Omega}_1$ is diagonal. In particular, each of its diagonal elements $\omega_{11}, \ldots \omega_{1F}$ has the following inverse gamma full conditional distribution

$$\omega_{1l}|\mathbf{Y}, \mathbf{X} \sim IG\left(\frac{1}{2}(n_1 + n_\omega), \frac{1}{2}\left(\sum_{i \in C_1}(x_{il} - \mu_{1l})^2 + n_{\omega l}s_{\omega l}^2\right)\right), \tag{3.6}$$

where x_{il} is the lth element of $\mathbf{x}_i$, μ_{1l} is the lth element of $\boldsymbol{\mu}_1$, and C_1 is the indicator of the first cluster.

The covariance matrix $\boldsymbol{\Omega}_k$ of the kth cluster, $k = 2, \ldots, K$, has the following inverse Wishart full conditional distribution

$$\boldsymbol{\Omega}_k|\mathbf{Y}, \mathbf{X} \sim IW\left(n_k + \nu, \sum_{i \in C_k}^{n_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)' + \boldsymbol{\Psi}_k\right), \tag{3.7}$$

where n_k is the size of the kth cluster and C_k is the indicator of the kth cluster.

To simulate the matrix of factor loadings $\mathbf{B}$, we separate two cases when $r > F$ and $1 < r \leq F$. Let $\mathbf{B}_r$ be the rth row of $\mathbf{B}$. For $r > F$, the full conditional distribution of $\mathbf{B}_r$ is the multivariate Gaussian distribution

$$\mathbf{B}_r|\mathbf{Y}, \mathbf{X} \sim N\left(\left(\frac{1}{\sigma_r^2}\mathbf{X}'\mathbf{X} + \mathrm{T}^{-1}\right)^{-1} \frac{1}{\sigma_r^2}\mathbf{X}'\mathbf{y}_{.,r}, \left(\frac{1}{\sigma_r^2}\mathbf{X}'\mathbf{X} + \mathrm{T}^{-1}\right)^{-1}\right), \tag{3.8}$$

where $\mathbf{y}_{.,r}$ is the rth column of $\mathbf{Y}$. Now let us consider the case when $1 < r \leq F$. Due to the hierarchical structural constraint, the last $F - r + 1$ elements of $\mathbf{B}_r$ are fixed. Thus, for $1 < r \leq F$, there are $r - 1$ free elements in $\mathbf{B}_r$, and no free elements when $r = 1$. Let $\mathbf{X}_{.,1:(r-1)}$ be submatrix of $\mathbf{X}$ before the rth column and $\mathbf{X}_{.,r}$ be the rth column of $\mathbf{X}$. Also, let $\mathrm{T}_{1:(r-1),1:(r-1)}$ be the submatrix of the first $r - 1$ rows and columns of the factor loading covariance matrix T. Then the full conditional is the Gaussian distribution $\mathbf{B}_r \sim N(\mathbf{Q}_r\mathbf{a}_r, \mathbf{Q}_r)$ where

$$\begin{aligned} \mathbf{Q}_r &= \left(\frac{1}{\sigma_r^2}\left(\mathbf{X}'_{.,1:(r-1)}\mathbf{X}_{.,1:(r-1)}\right) + \mathrm{T}_{1:(r-1),1:(r-1)}\right)^{-1}, \\ \mathbf{a}_r &= \frac{1}{\sigma_r^2}\left(\mathbf{X}'_{.,1:(r-1)}(\mathbf{y}_{.,r} - \mathbf{X}_{.,r})\right). \end{aligned} \tag{3.9}$$

The full conditional distribution of the rth idiosyncratic variance $\sigma_r^2, r = 1, \ldots, R$, is the inverse gamma distribution

$$\sigma_r^2|\mathbf{Y}, \mathbf{X} \sim IG\left(\frac{1}{2}(n_\sigma + S), \frac{1}{2}\left(n_\sigma s_\sigma^2 + \sum_{i=1}^{n}(\mathbf{Y}_{ir} - \mathbf{B}_r\mathbf{x}_i)^2\right)\right). \tag{3.10}$$

The full conditional distribution of the variance among factor loadings of the lth factor $\tau_l, l = 1, \ldots F$, is the inverse gamma distribution. Recall that the first l elements of the lth factor are fixed at 0 or 1 by the hierarchical structural constraint. Let $\mathbf{B}_{(l+1:r),l}$ be the lth factor after the first l elements. Then, the full conditional distribution of τ_l is

$$\tau_l|\mathbf{B}_{(l+1):r,l} \sim IG\left(\frac{1}{2}(R - l + n_\tau), \frac{1}{2}(\mathbf{B}'_{(l+1):r,l}\mathbf{B}_{(l+1):r,l} + n_\tau s_\tau^2)\right). \tag{3.11}$$

The full conditional distribution of the cluster assignment of the ith subject z_i is the discrete distribution

$$p(z_i = k|\mathbf{Y}, \mathbf{X}) \propto p_k|\mathbf{\Omega}_k|^{-1/2} exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)'\mathbf{\Omega}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right) \tag{3.12}$$

The full conditional distribution of probabilities $p_1, \ldots, p_K$ is the Dirichlet distribution

$$p_1, \ldots, p_K|\mathbf{Y}, \mathbf{X} \sim Dirichlet(n_1 + \alpha_1, \ldots, n_K + \alpha_K). \tag{3.13}$$

Using the full conditional distributions, we simulate BCFM by the following steps.

1. Set initial values of $\mathbf{X}$, $\mathbf{B}$, $\sigma_1^2, \ldots, \sigma_R^2$, and $\tau_1, \ldots, \tau_F$.
2. Set initial values of $\boldsymbol{\mu}_k$ and $\mathbf{\Omega}_k$. To do this, we run principal component analysis, k-means clustering, and LDL decomposition as specified in Section 3.2.2.
3. Set initial values of $z_1, \ldots, z_n$ and $\mathbf{p}$.
4. Simulate $\mathbf{X}$ from the Gaussian distribution given in Equation (3.4).
5. Simulate $\boldsymbol{\mu}$ from the Gaussian distribution given in Equation (3.5).
6. Simulate each $\omega_{1l}, \ldots \omega_{1F}$ from the inverse gamma distribution given in Equation (3.6) and $\mathbf{\Omega}_2, \ldots, \mathbf{\Omega}_K$ from the inverse Wishart distribution given in (3.7).
7. Simulate $\mathbf{B}$ from the Gaussian distributions given in Equations (3.8) and (3.9).
8. Simulate each $\sigma_1^2, \ldots, \sigma_R^2$ from the inverse gamma distribution given in Equation (3.10).
9. Simulate each $\tau_1, \ldots \tau_F$ from the inverse gamma distribution given in Equation (3.11).
10. Simulate each $z_1, \ldots, z_n$ from the discrete distribution given in Equation (3.12).
11. Simulate $\mathbf{p}$ from the Dirichlet distribution given in Equation (3.13).
12. Repeat steps **4** - **11** until the MCMC algorithm converges and we have enough posterior draws.

3.3.2 Model Evaluation

Usually in practice, the number of clusters and number of factors are not known. In this section, we propose two model selection methods to choose the number of clusters and the number of factors. First, we suggest the Laplace-Metropolis estimator of the marginal density [34]. This estimator assumes that the posterior distribution of the parameters is approximately Gaussian. Let the vector of parameters be $\boldsymbol{\Theta} = (\mathbf{X}, \mathbf{B}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2, \mathbf{z}, \mathbf{p})$, where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_R^2)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_F)$ and $\mathbf{z} = (z_1, \ldots, z_n)$. Then the marginal data density is

$$\int p(\mathbf{y}|K, F, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}|K, F) d\boldsymbol{\Theta}. \tag{3.14}$$

We compute the Laplace-Metropolis approximation of Equation (3.14). For the integrated likelihood, we integrate out the common factors $\mathbf{X}$ and cluster assignments $\mathbf{z}$. Then, the logarithm of integrated likelihood of a K-cluster and F-factor model is

$$\begin{aligned} p(\mathbf{y}|K, F, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\mathbf{p}}) = & \prod_{i=1}^{n} \sum_{k=1}^{K} \widehat{p}_k (2\pi)^{-R/2} |\widehat{\mathbf{B}} \widehat{\boldsymbol{\Omega}}_k \widehat{\mathbf{B}}' + \widehat{\mathbf{V}}|^{-1/2} \\ & exp\left(-\frac{1}{2} (\mathbf{y}_i - \widehat{\mathbf{B}} \widehat{\boldsymbol{\mu}}_k)' (\widehat{\mathbf{B}} \widehat{\boldsymbol{\Omega}}_k \widehat{\mathbf{B}}' + \widehat{\mathbf{V}})^{-1} (\mathbf{y}_i - \widehat{\mathbf{B}} \widehat{\boldsymbol{\mu}}_k) \right), \end{aligned} \tag{3.15}$$

where $\widehat{\mathbf{B}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2$ and $\widehat{\mathbf{p}}$ are the posterior means of $\mathbf{B}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2$ and $\mathbf{p}$ computed from the output of the MCMC algorithm. Therefore, the Laplace-Metropolis estimator of the marginal density in BCFM is

$$(2\pi)^{d/2} |\boldsymbol{\Phi}|^{1/2} p(\mathbf{y}|K, F, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\mathbf{p}}) p(\widehat{\mathbf{B}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\mathbf{p}}|K, F), \tag{3.16}$$

where $\boldsymbol{\Phi}$ is the posterior covariance matrix of the MCMC sample of $\mathbf{B}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2$ and $\mathbf{p}$. Also, d is the dimension of $\boldsymbol{\Phi}$. Thus,

$$d = \frac{(K-2)(F+1)}{2} + (R+K)(F+1) + F - 1.$$

Another model selection criterion we propose to choose the number of clusters and the number of factors of BCFM is similar to the Bayesian Information Criterion (BIC). While

the original BIC considers the maximum likelihood estimates of the parameters, here we use the posterior means. Thus, the criterion we propose is defined as

$$\mathbf{BIC}_{like} = d\log(n) - 2\log p(\mathbf{y}|K, F, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\mathbf{p}}), \tag{3.17}$$

where $p(\mathbf{y}|K, F, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\mathbf{p}})$ is the integrated likelihood from Equation (3.15).

In what follows, we consider these two model selection criteria to select the number of clusters and the number of factors.

3.4 Simulation Studies

3.4.1 Evaluation of Estimation

To evaluate the quality of estimation, we consider a dataset simulated with $n = 1,000$ subjects, $R = 20$ variables, $K = 4$ clusters, and $F = 3$ factors. This simulation setting has been inspired by our recent analysis of individuals in recovery from opioid use disorder [14]. In our setting, the 1,000 subjects are randomly assigned to the 4 clusters with probabilities (0.45, 0.30, 0.15, 0.10). The true mean vectors of the common factors of each cluster are $\boldsymbol{\mu}_1 = (0.3, -0.8, -0.1)$, $\boldsymbol{\mu}_2 = (-3.0, -8.0, 5.0)$, $\boldsymbol{\mu}_3 = (-7.5, 5.0, 2.0)$, and $\boldsymbol{\mu}_4 = (-15.0, -3.5, 10.5)$. The true values of the covariance matrices of the common factors of each cluster are

$$\boldsymbol{\Omega}_1 = \begin{bmatrix} 1.88 & 0.00 & 0.00 \\ 0.00 & 1.08 & 0.00 \\ 0.00 & 0.00 & 1.33 \end{bmatrix}, \boldsymbol{\Omega}_2 = \begin{bmatrix} 2.00 & 0.40 & 0.40 \\ 0.40 & 2.00 & 0.40 \\ 0.40 & 0.40 & 2.00 \end{bmatrix},$$

$$\boldsymbol{\Omega}_3 = \begin{bmatrix} 3.00 & 0.45 & 0.45 \\ 0.45 & 3.00 & 0.45 \\ 0.45 & 0.45 & 3.00 \end{bmatrix}, \text{ and } \boldsymbol{\Omega}_4 = \begin{bmatrix} 4.00 & 1.00 & 1.00 \\ 1.00 & 4.00 & 1.00 \\ 1.00 & 1.00 & 4.00 \end{bmatrix}.$$

To simulate the matrix of factor loadings, the variances of the free elements in each of its columns were $\tau_1 = 0.05$, $\tau_2 = 0.10$, and $\tau_3 = 0.15$. In addition, the idiosyncratic variances

$\sigma_j^2 = 0.1$, for $j = 1, \ldots, R$.

To analyze the simulated dataset, first we set up the priors as explained in Section 3.2.2. After that, we run the MCMC algorithm proposed in Section 3.3.1 for 50,000 iterations. To reduce computer memory burden we have kept one draw for each 10 iterations, retaining a total of 5,000 draws. Trace plots of the simulated quantities indicate that the algorithm converged after 1,500 draws (i.e., 15,000 iterations.) Thus, we discard the first 1,500 draws as burn-in and use the remaining 3,500 draws to estimate the BCFM parameters. The following figures in this section present the results when fitting the model with $K = 4$ clusters and $F = 3$ factors.

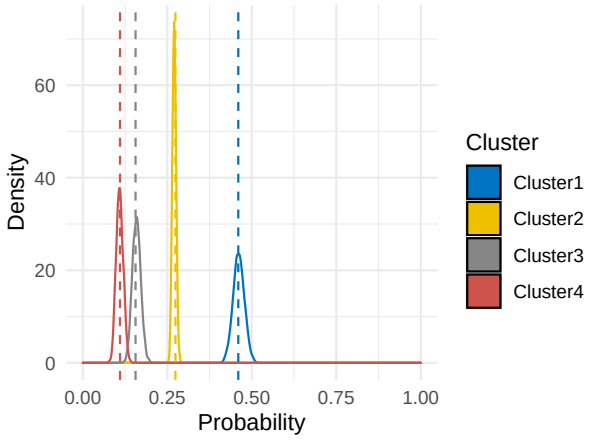


Figure 3.1: Simulated data – posterior densities (solid lines) of the cluster probabilities for BCFM with $K = 4$ clusters and $F = 3$ factors. For comparison, vertical dashed lines indicate the true values of the cluster probabilities.

Figure 3.1 shows the posterior density of each of the cluster probabilities $\mathrm{p}_1, \mathrm{p}_2, \mathrm{p}_3$, and p_4, as well as their respective true values (vertical dashed lines). The posterior modes are close to the true values and, thus, our BCFM framework is able to accurately estimate these probabilities. In addition, all true cluster probabilities fall within the respective 95% credible intervals for each cluster. Therefore, our BCFM framework provides adequate quantification of uncertainty for the cluster probabilities.

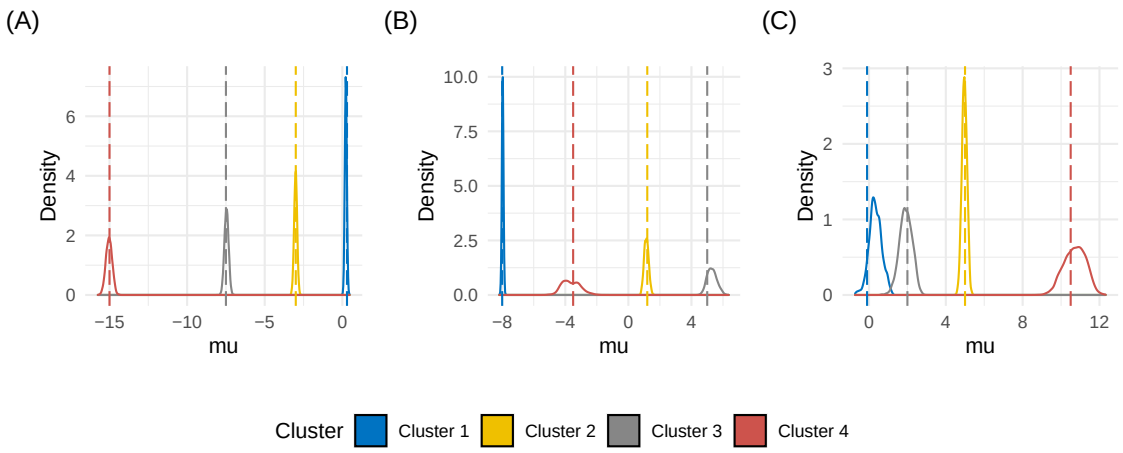


Figure 3.2: Simulated data – posterior densities of the elements of the mean vectors for the common factors of each cluster. (A–C) Each panel corresponds to the means of a common factor across clusters. (A) first factor, (B) second factor, and (C) third factor. For comparison, vertical dashed lines indicate the true values of the means of the common factors within each cluster.

Figure 3.2 displays the posterior densities of the cluster means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_4$, where each of these vectors contains three elements, one for each factor. Panel (A) shows the posterior densities of the first element — which corresponds to the mean of the first common factor — of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_4$. Panels (B) and (C) show analogous plots for the second and third elements, respectively. The modes of the posterior distributions are very close to the true values, the posterior densities are highly concentrated, and the true values are located in regions of high posterior mass. Therefore, our BCFM framework provides accurate estimates and adequate quantification of uncertainty for the cluster means.

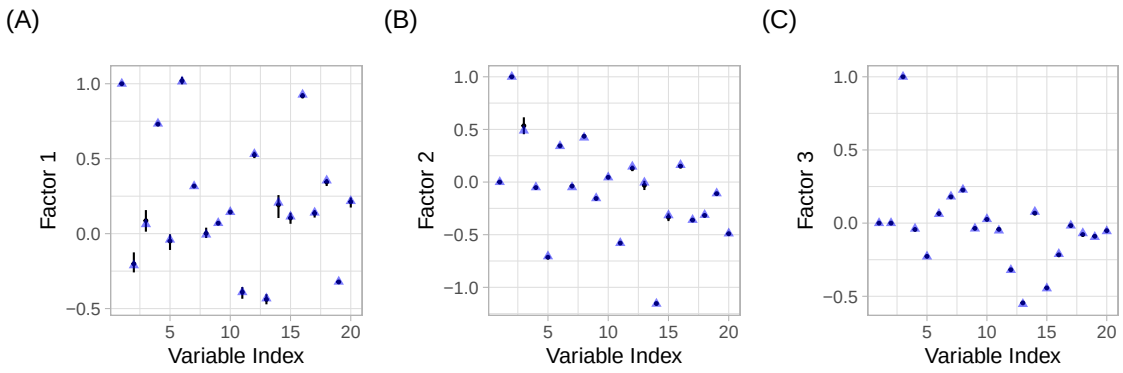


Figure 3.3: Simulated data – posterior summaries of factor loadings for BCFM with $K = 4$ clusters and $F = 3$ factors: true value (blue triangle), posterior mean (black circle), and 95% credible interval (black vertical line). (A) first factor, (B) second factor, and (C) third factor.

Figure 3.3 shows the posterior mean and 95% credible intervals of the factor loadings. Note that because of the hierarchical structural constraint, for the lth factor, the lth loading is fixed at 1 and the loadings in positions at 1 to $l-1$ are fixed at 0. The credible intervals are very narrow, and the posterior means overlap the true values. True values are represented with blue triangles, posterior means are represented with black circles, and black vertical lines indicate 95% credible intervals. Our BCFM framework accurately estimates the factor loadings. In addition, 98% of the true factor loadings are included in the 95% credible intervals, indicating an appropriate quantification of uncertainty.

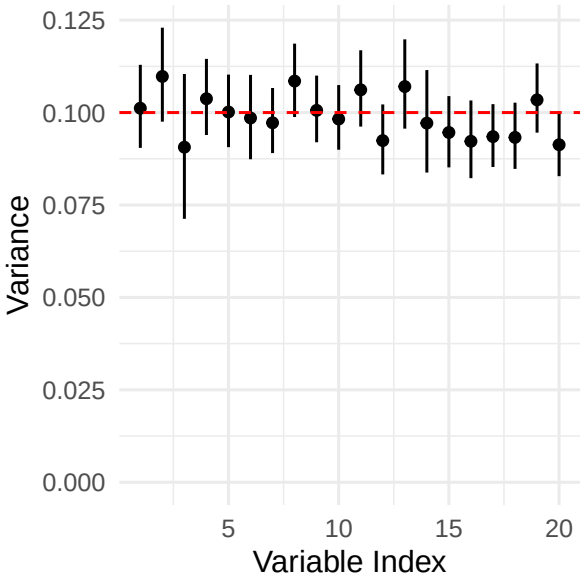


Figure 3.4: Simulated data – idiosyncratic variances for BCFM with $K = 4$ clusters and $F = 3$ factors: 95% credible interval (vertical line) and posterior mean (circle), and true values (red dashed line).

Figure 3.4 shows true values (red dashed line), posterior means (black circles), and 95% credible intervals (black vertical lines) for σ_r^2, $r = 1, \ldots, 20$. Recall that the true values used to generate the simulated data are $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{20}^2 = 0.1$. The posterior means are close to the true values indicating that our approach accurately estimates the idiosyncratic variances. In addition, the 95% credible intervals include the true values 100% of the time. Therefore, our approach appropriately quantifies the uncertainty in the estimation of the idiosyncratic variances.

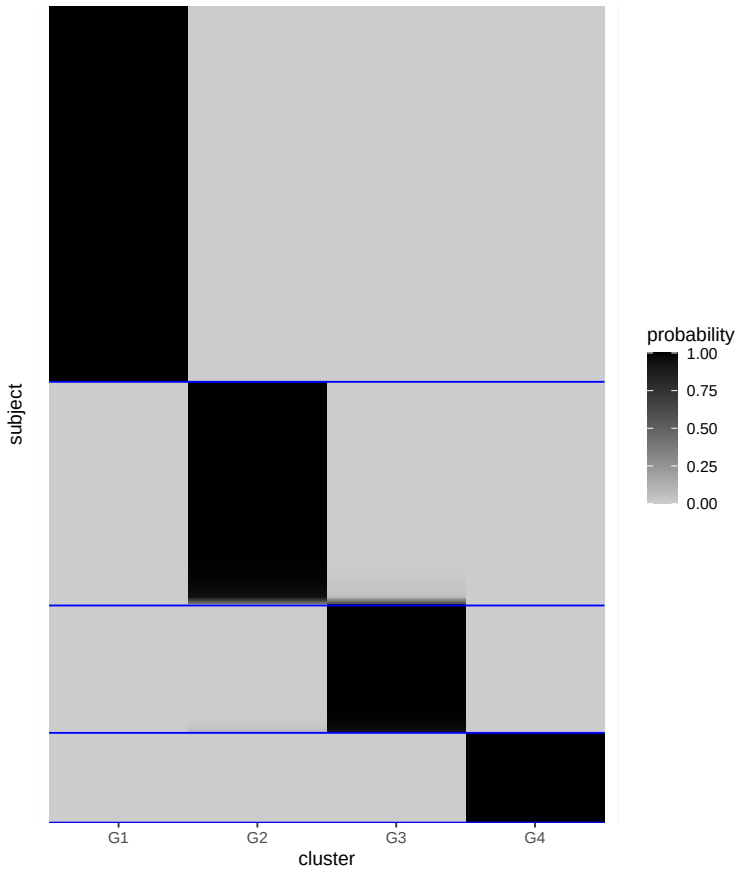


Figure 3.5: Simulated data – heatmap of the cluster assignment probabilities and the true clusters for BCFM with $K = 4$ clusters and $F = 3$ factors. Blue lines present the boundaries of the true clusters. The shades represent the posterior probability that each subject belongs to each cluster.

Figure 3.5 presents the heatmap of the posterior probability that each subject belongs to each cluster. The x-axis represents the 4 clusters, and the y-axis represents the 1,000 subjects. The blue lines separate the true clusters. Our BCFM framework correctly assigns 99.6% of the subjects to their true clusters. Therefore, our approach assigns subjects to clusters with high accuracy.

Model	1 Cluster	2 Clusters	3 Clusters	4 Clusters	5 Clusters
1 Factor	-41829	-546869	-41197	-41218	-41610
2 Factors	-21863	-20967	-20667	-20597	-20615
3 Factors	-14953	-13670	-13222	**-13128**	-13157
4 Factors	-14952	-13675	-13226	-13144	-13173
5 Factors	-14949	-13680	-13243	-13171	-13214

Table 3.1: Simulated dataset – Laplace-Metropolis estimator of the marginal density

Model	1 Cluster	2 Clusters	3 Clusters	4 Clusters	5 Clusters
1 Factor	83494	1091799	82231	82264	82917
2 Factors	43548	41762	41174	41047	41095
3 Factors	29684	27137	26264	**26098**	26174
4 Factors	29803	27283	26451	26314	26419
5 Factors	29928	27449	26650	26550	2669

Table 3.2: Simulated dataset – BIC with integrated likelihood criterion

Table 3.1 shows the result of the Laplace-Metropolis marginal density of the marginal density, and Table 3.2 shows the result of the BIC with the integrated likelihood. The largest value in Table 3.1 is when using a 4-cluster and 3-factor model, which is the true setting of this simulated dataset. The smallest value in Table 3.2 also comes from 4 clusters and 3 factors. Thus, two criteria agree the result of the true model is preferable.

3.4.2 Simulation Study for Model Selection

To further validate our BCFM framework, we perform a simulation study to compare the BCFM model selection approach with the method proposed by [44]. Specifically, [44] proposes a method based on overfitting a Bayesian mixture model which is implemented in the R package `fabMix`, henceforth referred to as the fabMix method. Here, we compare BCFM and fabMix in terms of the performance of correctly choosing the number of clusters and

factors. In particular, we implemented `fabMix` using the following parameter values: (1) `model = 'UCU'`, which is equivalent to the model specification for BCFM; (2) `nchains = 2`; (3) `mCycles = 1000`; (4) `burnCycles = 100`; (5) `q=c(1,2,3,4,5)`; (6) `nIterPerCycle = 10`; and (7) all other parameters were set to the default options.

We have simulated 100 datasets with $K = 4$ clusters and $F = 3$ factors for each of three settings. The settings are when the clusters are well-separated, moderately separated, and not-well-separated. The well-separated setting has cluster mean vectors $\boldsymbol{\mu}_1 = (0, -4, 4)$, $\boldsymbol{\mu}_2 = (6, 2, -2)$, $\boldsymbol{\mu}_3 = (-4, 10, 8)$, and $\boldsymbol{\mu}_4 = (-10, -12, 14)$. We have halved these vectors to obtain the cluster mean vectors for the moderately separated setting which are $\boldsymbol{\mu}_1 = (0, -2, 2)$, $\boldsymbol{\mu}_2 = (3, 1, -1)$, $\boldsymbol{\mu}_3 = (-2, 5, 4)$, and $\boldsymbol{\mu}_4 = (-5, -6, 7)$. We again halved these vectors to obtain the cluster mean vectors for the not-well-separated dataset, which are $\boldsymbol{\mu}_1 = (0, -1, 2)$, $\boldsymbol{\mu}_2 = (2, -1, 0)$, $\boldsymbol{\mu}_3 = (-1.5, 3.5, 3)$, and $\boldsymbol{\mu}_4 = (-3, -2.5, -1.5)$.

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	0	0
F=3	0	0	0	6	2
F=4	0	0	0	45	5
F=5	0	0	0	42	0

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	0	0
F=3	0	0	0	99	1
F=4	0	0	0	0	0
F=5	0	0	0	0	0

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	0	0
F=3	0	0	0	100	0
F=4	0	0	0	0	0
F=5	0	0	0	0	0

Table 3.3: Marginal density (top-left), BIC with integrated likelihood criterion (top-right) and `fabMix` (bottom) result of the 100 well separated datasets

Table 3.3 shows the number of best model selections of the 100 well separated datasets. The BIC with integrated likelihood criterion chose the correct model 99 times and concluded once that the most adequate model has 5 clusters and 3 factors. The Laplace-Metropolis

estimator of the marginal density was able to find the right number of clusters 93 times but often overestimated the number of factors. The `fabMix` found the correct number of clusters and factors every time in the well separated case.

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	1	0
F=3	0	0	0	0	0
F=4	0	0	27	50	21
F=5	0	0	0	1	0

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	0	0
F=3	0	1	0	99	0
F=4	0	0	0	0	0
F=5	0	0	0	0	0

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	0	0
F=3	0	18	14	36	0
F=4	0	22	0	0	0
F=5	0	5	0	0	0

Table 3.4: Marginal density (top-left), BIC with integrated likelihood criterion (top-right) and `fabMix` (bottom) of the 100 moderately separated datasets

Table 3.4 shows the number of best model selections of the 100 moderately separated datasets. The BIC with integrated likelihood criterion concluded one time with 2 clusters and 3 factors, but all the other models with the true setting. The marginal density did not find the true number of clusters and the factors simultaneously but found the correct number of clusters 98% of the time. Our competing method from `fabMix` found the correct number of clusters and factors 36 times.

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	1	0
F=3	0	8	0	0	0
F=4	7	23	0	0	0
F=5	46	16	0	0	0

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	0	0
F=3	26	74	0	0	0
F=4	0	0	0	0	0
F=5	0	0	0	0	0

Model	K=1	K=2	K=3	K=4	K=5
F=1	0	0	0	0	0
F=2	0	0	0	0	0
F=3	99	0	0	0	0
F=4	0	0	0	0	0
F=5	0	1	0	0	0

Table 3.5: Marginal density (top-left), BIC with integrated likelihood criterion (top-right) and `fabMix` (bottom) of the 100 slightly separated datasets

Table 3.5 shows the number of best model selections from the 100 slightly separated datasets. The BIC with integrated likelihood criterion found the true number of factors every time. It selected 74 datasets with 2 clusters and the rest with no cluster. The marginal density often overestimated the number of factors. Also, it found 47% of the datasets with 2 clusters and the others without clusters. At last, `fabMix` package found the right number of factors 99 times, but without any clusters.

From the results, we found that the BIC with integrated likelihood criterion mostly selects the correct number of factors and often the true number of clusters. The `fabMix` package found the true settings correctly in the well separated case. However, when the distances among clusters are smaller, it often underestimates the number of clusters. The Laplace-Metropolis estimator of the marginal density often overestimates the number of factors. From this result, we can find BCFM with the BIC with integrated likelihood criterion provides better results.

3.5 Applications

This section presents the application of BCFM to two real datasets. First, we present an analysis of a dataset on opioid use disorder (OUD). Second, we apply BCFM to a dataset on breast cancer gene expression.

3.5.1 Opioid Use Disorder Recovery Data

This section presents a BCFM analysis of a dataset on recovery from opioid use disorder (OUD). The data are from the Remission from Chronic Opioid Use – Studying Environmental and SocioEconomic Factors on Recovery (RECOVER, NCT03604861) Study [**?**]. While the RECOVER study collected data for 24 months, here we only analyze the data from the first time point. In the data we consider, there are $n = 530$ participants and $R = 8$ variables. The variables are the Subjective Opiate Withdrawal Scale (SOWS), Beck's Depression Inventory II (BDI), Family & Social score, average pain score of Brief Pain Inventory (BPI), one question about the need for lifetime OUD medication, one question about confidence in abstinence, and the physical and mental categories of the 12-Item Short Form Survey (SF-12).

Model	1 Cluster	2 Clusters	3 Clusters	4 Clusters	5 Clusters
1 Factor	-5680	-5471	-5402	-5422	-5431
2 Factors	-5634	-5383	-5313	-5280	-5261
3 Factors	-5611	-5111	-4752	**-4217**	-4243
4 Factors	-5614	-4829	-4704	-4571	-4324
5 Factors	-5642	-4740	-4614	-4918	-4495

Table 3.6: OUD Recovery data – Laplace-Metropolis estimator of the marginal density.

Model	1 Cluster	2 Clusters	3 Clusters	4 Clusters	5 Clusters
1 Factor	11344	10838	10677	10707	10719
2 Factors	11253	10674	10488	10450	10424
3 Factors	11254	9774	8933	**7568**	7647
4 Factors	11291	9721	9049	8975	7835
5 Factors	11335	9426	9099	9788	9108

Table 3.7: OUD Recovery data – BIC with integrated likelihood criterion.

We evaluate the result with two criteria introduced in Section 3.3.2. Table 3.6 presents the result of the marginal density. The model with 4 clusters and 3 factors had the largest value. Table 3.7 shows the BIC with integrated likelihood criterion of the OUD Recovery dataset. According to this method, the 4-cluster and 3-factor model had the largest value. Therefore, both evaluation methods agree that the 4-cluster and 3-factor model performs best. The figures in the rest of this section are the parameters from the posterior of this model.

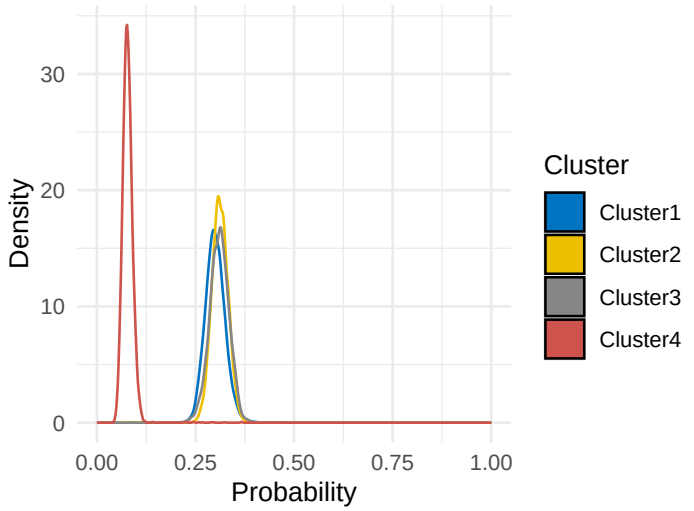


Figure 3.6: OUD Recovery data – density of group assignment probabilities.

Figure 3.6 shows the posterior density of the cluster assignment probabilities. The three clusters are about the same size, while the other cluster is much smaller. The posterior mean of the assignment probabilities is (0.31, 0.31, 0.3, 0.08).

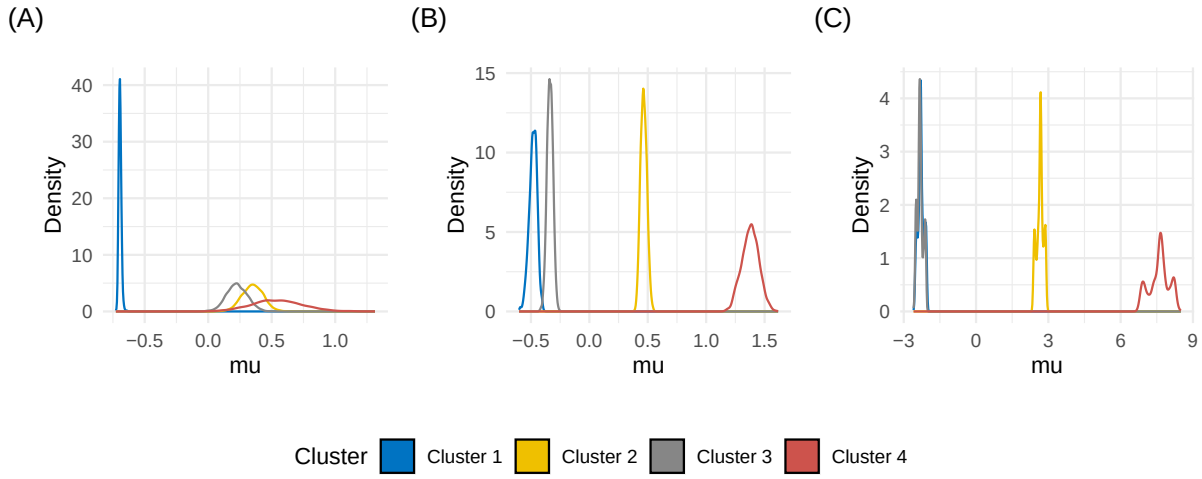


Figure 3.7: OUD Recovery data – posterior density of the cluster means. (A – C) Each panel corresponds to one of the dimensions of the estimated posterior means. The colors represent different clusters.

Figure 3.7 shows the posterior density of the cluster means. The first cluster (black) has a smaller variance than the other clusters. The fourth cluster (red), which has the smallest cluster probability, has the largest variance in all three factors.

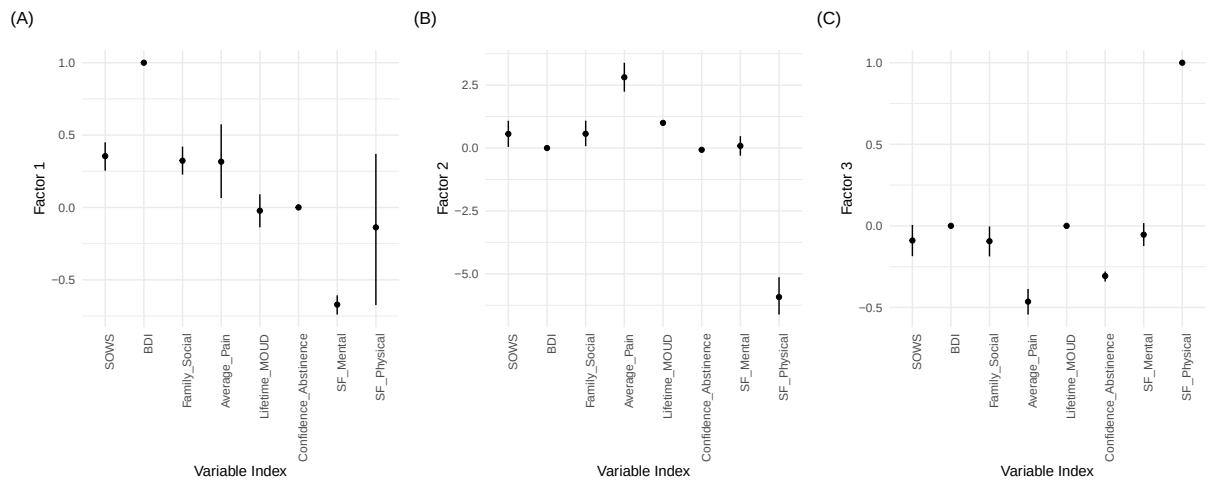


Figure 3.8: OUD Recovery data – posterior density of the factor loadings. Posterior mean (black circle) and 95% credible intervals (black line).

Figure 3.8 shows the posterior density of the factor loadings. The second variable, BDI, is fixed at 1 in the first factor. The fifth variable, the question asking about the need for lifetime medication, is fixed at 1 in the second factor. Compared to this value, the factor loading of the eighth variable, the mental score of SF-12, has the largest factor loading and uncertainty. This variable was selected to be fixed at 1 in the third factor. The posterior mean of the other variables in the third factor have a much smaller size than 1.

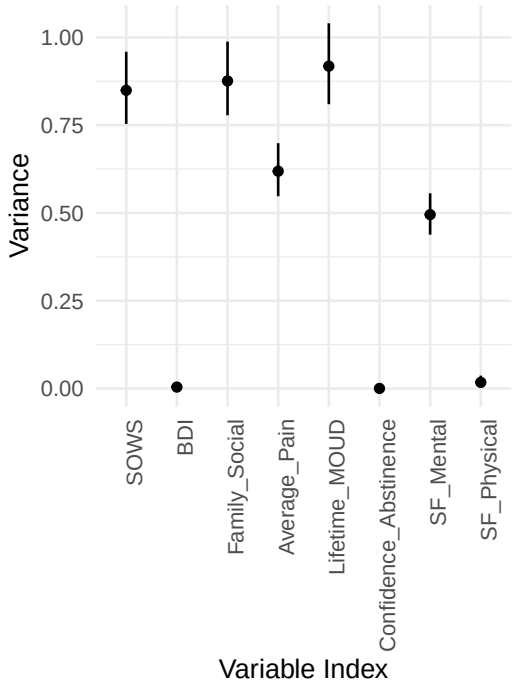


Figure 3.9: OUD Recovery data – posterior density of the idiosyncratic error variance, $\boldsymbol{\sigma}^2$. Posterior mean (black circle), 95% credible intervals (black line) and value 1 (red dashed line.)

Figure 3.9 displays the posterior density of the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_8^2$. Idiosyncratic error variance of the second, sixth, and eighth variables are close to zero. The 95% credible intervals are also narrower than the other variables. The variable with the largest posterior mean is the fifth variable, and the value is about 0.9.

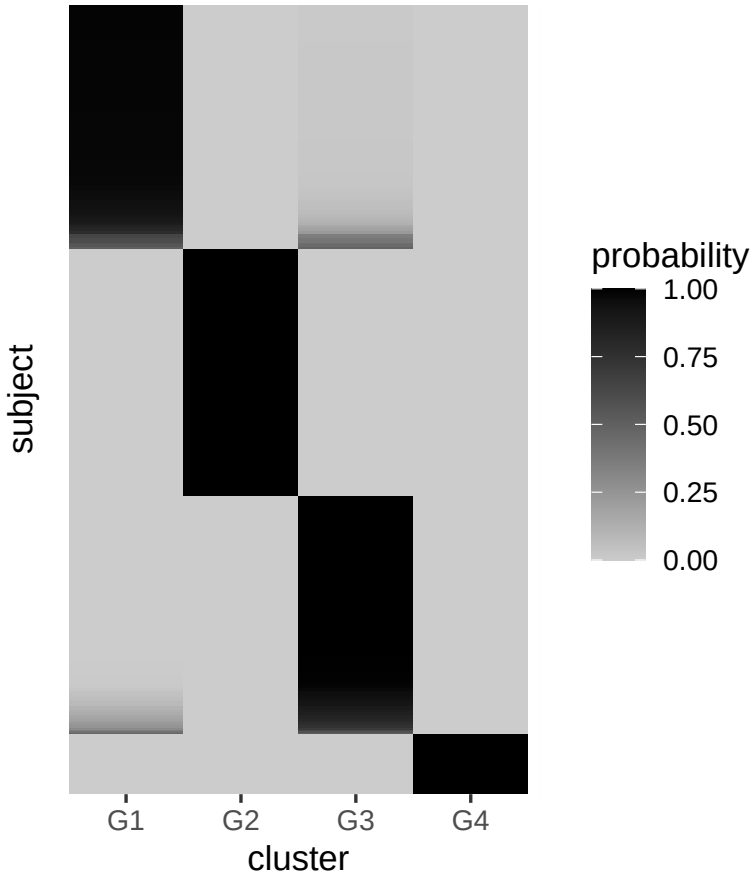


Figure 3.10: OUD Recovery data – heatmap of the cluster assignments. The subjects are ordered according to the largest cluster probabilities.

Figure 3.10 is the heatmap of the cluster assignments. Some observations may be confused with the first cluster and the third cluster during the MCMC. However, there is little variability when assigning to the second, third, and fourth clusters.

3.5.2 Breast Cancer Molecular Subtype Data

We illustrate the result of BCFM through the gene expression data from the breast cancer molecular subtypes. There are $n = 1,428$ patients, $R = 93$ gene expressions, and $K = 4$ breast cancer molecular subtypes: Luminal A, Luminal B, HER2-enriched, and Triple-negative. Luminal A is the majority type in this data (66.1%), and triple-negative took second place (20.9%). HER2-enriched breast cancer takes 8.9% of the data, while Luminal B has the least number of cases (4.1%).

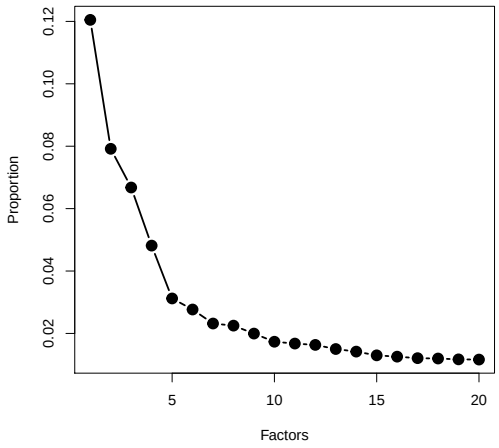


Figure 3.11: Breast Cancer data – variability explained through PCA. Number of factors (x-axis), proportion of variability explained (y-axis).

Figure 3.11 shows the proportion of variability the principal components explain. The first factor can explain about 12% of the data, and the factors after the first factor cannot account for less than 10% of the variability. With the first 20 factors, PCA can explain around 59% of the variability. From this information, we run models up to 6 clusters and 20 factors.

Model	1 Cluster	2 Clusters	3 Clusters	4 Clusters	5 Clusters	6 Clusters
1 Factor	-182129	-182115	-182227	-182351	-182322	-183011
2 Factors	-177633	-177530	-177515	-177510	-177572	-177670
3 Factors	-173530	-173308	-173202	-173182	-173196	-173190
4 Factors	-170750	-170399	-170286	-170146	-170127	-170141
5 Factors	-169534	-168977	-168833	-168706	-168670	-168657
6 Factors	-169562	-166940	-167672	-167569	-167477	-167981
7 Factors	-167783	-166267	-166727	-166609	-166631	-166633
8 Factors	-166948	-166267	-165897	-166040	-165889	-165854
9 Factors	-166437	-165654	-165412	-165282	-165408	-165225
10 Factors	-166244	-165480	-164999	-164841	-164914	-164847
11 Factors	-165952	-166232	-164560	-164543	-164443	-164308
12 Factors	-165507	-164351	-163855	-163642	-163689	-164093
13 Factors	-165229	-163974	-164051	-163802	-163292	-164232
14 Factors	-165035	-163704	-163335	-163114	-163140	-163069
15 Factors	-164939	-163641	-163023	-162917	-162919	-163012
16 Factors	-164782	-163316	-162850	-162840	-162783	-162932
17 Factors	-163903	-163394	-163051	-162602	-162738	-163062
18 Factors	-164724	-163167	-163782	**-162534**	-162813	-162932
19 Factors	-164604	-163081	-162678	-162741	-162767	-162685
20 Factors	-164601	-162989	-163178	-163346	-162642	-163634

Table 3.8: Breast Cancer data – Laplace-Metropolis estimator of the marginal density

Model	1 Cluster	2 Clusters	3 Clusters	4 Clusters	5 Clusters	6 Clusters
1 Factor	364016	363991	364044	364127	364086	365605
2 Factors	355175	354971	354930	354940	354983	355077
3 Factors	347118	346637	340826	346479	3465022	346534
4 Factors	341722	340989	338170	340574	340607	340664
5 Factors	339512	338320	338170	337909	337926	337963
6 Factors	337837	336353	336013	335916	335897	336352
7 Factors	335794	334515	334251	334166	334244	334387
8 Factors	334990	333254	332878	333273	333382	333221
9 Factors	334248	332584	332375	332272	332912	332768
10 Factors	333755	332443	331870	331645	331999	332240
11 Factors	333256	333138	331174	331441	331487	331655
12 Factors	333025	330530	329958	330042	330563	331554
13 Factors	332896	330222	330832	330819	330308	332411
14 Factors	332765	329915	329449	329786	330361	330660
15 Factors	332732	330306	329434	**329643**	330340	331068
16 Factors	332714	329887	329314	330112	330490	331422
17 Factors	332778	330541	330240	329881	331149	332065
18 Factors	332931	330050	332443	330244	331766	332413
19 Factors	333300	330102	330224	331309	332102	332751
20 Factors	333585	330458	331798	332924	332470	334966

Table 3.9: Breast Cancer data – BIC with integrated likelihood criterion

Table 3.8 shows the marginal density, and Table 3.9 shows the BIC with integrated likelihood criterion of the breast cancer dataset. According to the marginal density, 4 clusters and 18 factors BCFM has the best performance. On the other hand, the BIC with integrated likelihood criterion chose the model with 4 clusters and 15 factors. Both criteria agree on the correct number of clusters but different number of factors. Since the BIC with integrated likelihood criterion had better performance than the marginal density in Section **??**, we select

the result the model with fewer factors.

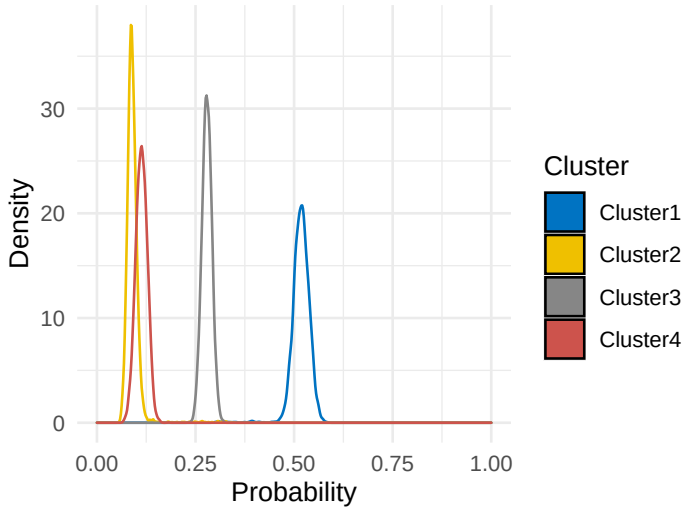


Figure 3.12: Breast Cancer data – posterior density of the probabilities. Posterior density (curves) and the true probabilities (dashed lines.)

Figure 3.12 displays the posterior density of the probabilities. About half the observations are assigned to the first cluster. This number is smaller than the number of patients diagnosed with Luminal A, the most common subtype in the dataset. In contrast, the smallest cluster in the 4-cluster-15-factor BCFM is larger than the size of the true group with the least case, molecular subtype Luminal B.

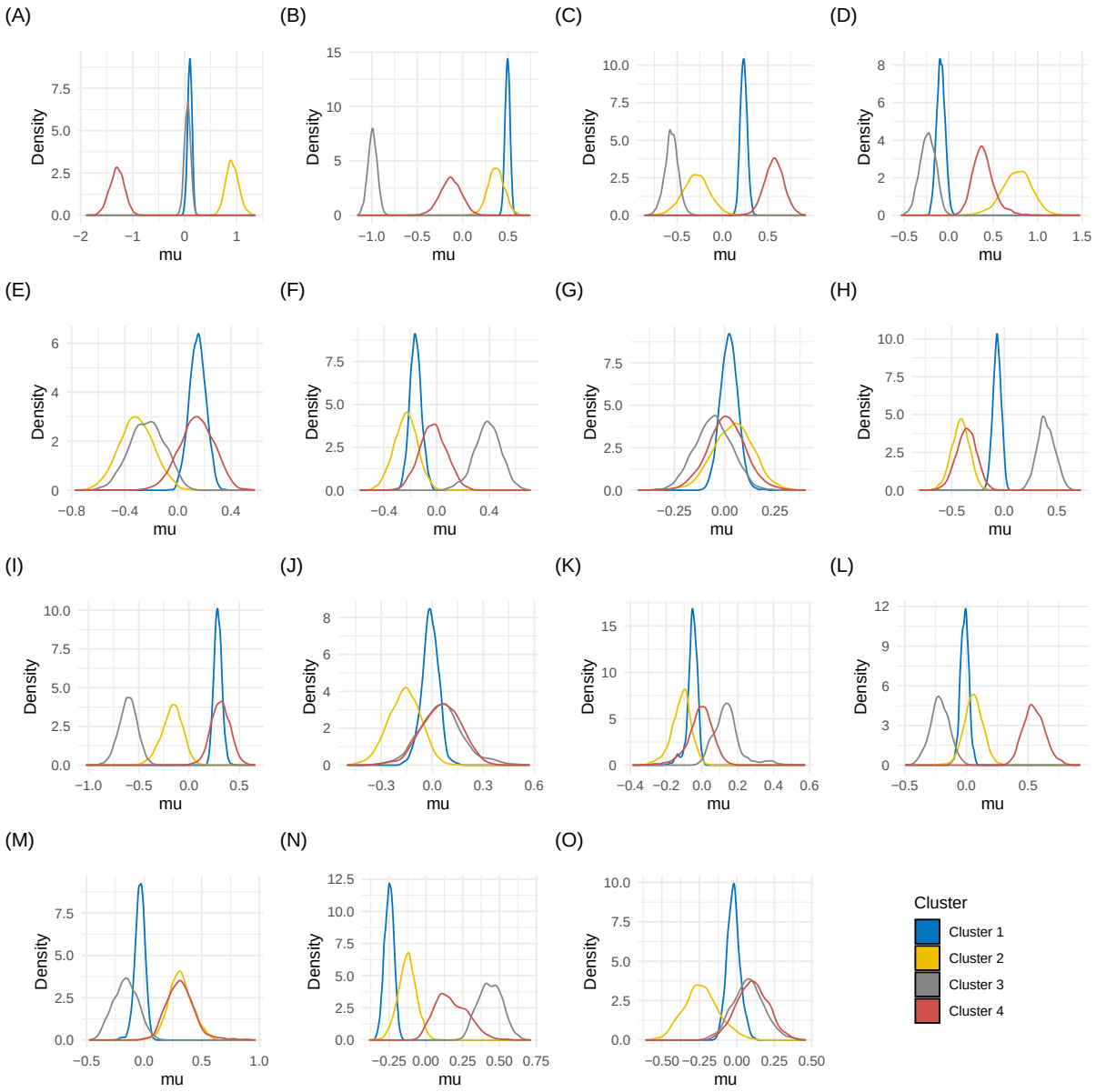


Figure 3.13: Breast Cancer data – posterior density of $\boldsymbol{\mu}$. (A–O) Each panel corresponds to one of the dimensions of the estimated posterior means. The colors encode the different clusters.

Figure 3.13 is the posterior density of the cluster means, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_4$. Each plot (A) - (O) represents the factor. There are factors where the posterior distribution of the four clusters represents different values and similar values. For example, in the seventh factor (G), the posterior modes of the four clusters are not well separated. In the third factor (C), the posterior modes of the clusters are distant.

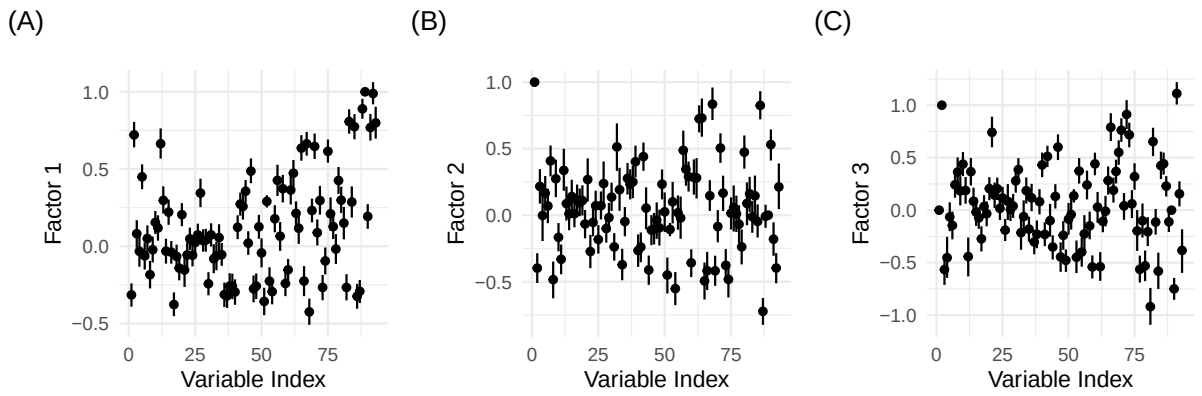


Figure 3.14: Breast Cancer data – posterior density of the factor loadings. Posterior mean (black circle), 95% credible interval (black line), and factor loadings fixed at 1 (red dashed line).

Figure 3.14 displays the posterior mean and the 95% credible intervals of the factor loadings of the first three factors. The factor loadings in the first three factors range between -1 to 1 and have small credible intervals.

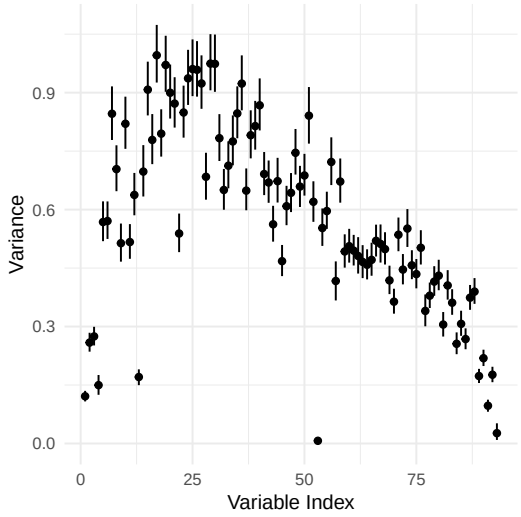


Figure 3.15: Breast Cancer data – posterior density of the idiosyncratic variance, σ^2. Posterior mean (black circle), 95% credible interval (black line) and value 1 (red dashed line.)

Figure 3.15 displays the posterior mean and 95% credible intervals of the idiosyncratic variances from the breast cancer dataset. The largest variance is around 1, while the smallest is close to 0. The variables with larger posterior mean usually have wider credible intervals.

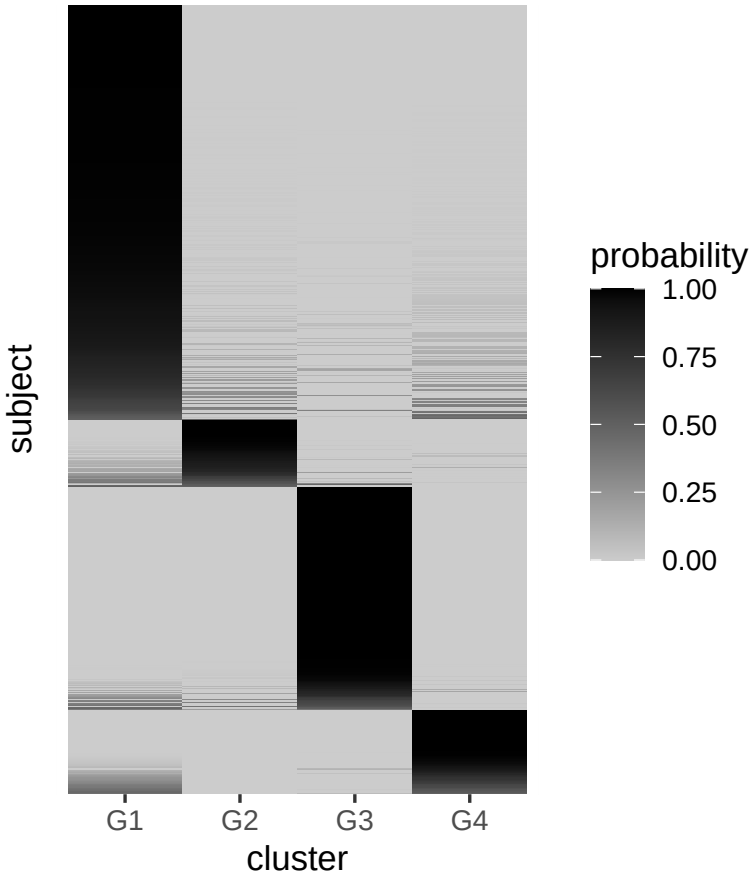


Figure 3.16: Breast Cancer data – heatmap of the BCFM cluster assignment probabilities and the true clusters. Blue lines present the boundaries of the true clusters. The shades represent the probability an observation is assigned to each cluster during the MCMC.

Figure 3.16 shows the heatmap of the BCFM cluster assignment probabilities and the true clusters. More than half of the patients with breast cancer type Luminal-A are assigned to the correct cluster. However, the BCFM may confuse that some of the patients belong to either the second or the fourth cluster. Our 4 clusters 15 factors BCFM found molecular subtype triple negative as the third cluster. Finally, the model cannot find the HER2-enriched and Luminal-B subtypes well. HER2-enriched is often assigned to the third cluster, and Luminal-B is mostly assigned to the first cluster.

3.6 Conclusions

In the Bayesian Clustering Factor Models, we incorporate the Gaussian mixture model on Bayesian factors. Our model finds the latent clusters and factors given the prespecified settings. The number of factors is usually smaller than the dimension of the data and therefore may achieve dimension reduction. For identifiability and interpretability of the model, we need two constraints on the model. We put hierarchical structural constraint on factor loadings and restrict the first cluster covariance matrix to be diagonal. We use informative priors on the cluster means, covariances, and assignment probabilities. Therefore, we have to choose the hyperparameters carefully. We apply principal component analysis and k-means clustering before BCFM to find the appropriate cluster hyperparameters. We proposed a Gibbs sampler to explore the posterior. We also developed two evaluation criteria: the Laplace-Metropolis estimator of the marginal density and the BIC with integrated likelihood criterion.

We presented the work of BCFM through simulated datasets, the OUD dataset, and the breast cancer molecular subtype dataset. The results of the single simulated dataset proved that our BCFM and the evaluation methods found the correct setting of the data. We also generated 300 datasets for 3 different options, where the datasets are well separated, moderately separated, and slightly separated. We compared our method with the overfitting Bayesian mixture model from [44]. The performance level was similar in the well separated datasets, but BCFM with BIC with integrated likelihood criterion outperformed when applied to moderately separated and slightly separated datasets. In the OUD dataset, BCFM defined 4-cluster-3-factor as the best model. Each cluster displayed different behavior. In the breast cancer dataset, both evaluation methods agreed on the number of clusters, but the BIC with integrated likelihood criterion would select fewer factors than the Laplace-Metropolis estimator of the marginal density. The model selected by the BIC with integrated likelihood criterion had the size of the clusters and the cluster assignments different from the true setting. However, the best BCFM, according to both assessment methods, chose the correct number of clusters.

An interesting future avenue for this research would be to incorporate spatial and temporal structures. When applying BCFM to datasets with time series observations and geospatial information, our model cannot consider the possible correlations from those. Thus, research

that explains the variability coming from the spatiotemporal components would be effective. Another possible future study would be finding an evaluation criterion for BCFM. From the moderately separated and slightly separated simulated datasets, we found the BIC with integrated likelihood criterion would choose smaller models, and marginal density would select larger models. A new evaluation method that finds the correct setting even when the clusters are not well separated would be practical in this case. An innovative assessment method to find the settings would allow better selection of the BCFM.

Chapter 4

Packages

4.1 Introduction

In this chapter, we introduce two packages from Chapter 2 and Chapter 3, `DIFM` and `BCFM`. The packages provide functions to run MCMC to generate posterior sample and provide data visualization.

4.2 DIFM Package

Package `DIFM` was built with R 4.3.2 [48] and includes functions to explore data, run the Gibbs sampler, generate plots, and evaluate final models of Dynamic ICAR Spatiotemporal Factors Model. The gibbs sampler is implemented in C++ for faster computations.

The package depends on following packages: `Rcpp, RcppArmadillo, Matrix, LaplacesDemon, spdep, gridExtra` and `spdep`. Package `Rcpp` [17] and `RcppArmadillo` [18] are necessary to plug in C++ into R. Package `LaplacesDemon` [52] is used to simulate parameters from their full conditionals that follow inverse gamma and inverse Wishart distribution. We extract the neighborhood matrix for the ICAR prior and generate polygon maps through `sp` [46] and `spdep` [8] packages. At last, we use package `gridExtra` to collate spatial plots into a single window.

4.2.1 Documentation of DIFM Package

This package has 23 R functions, 7 C++ functions and 2 datasets. The functions that are accessible to users are 13 R functions and 1 C++ function. Below is the list of functions

that are available to users.

- `buildH`: Computes the spatial covariance and precision matrix of the neighboring subregions that is necessary for Intrinsic Autoregressive Conditional (ICAR) process.
  Usage: `buildH(areapoly, permutation)`

- `difm.hyp.parm`: Sets the hyperparameters of $\boldsymbol{\tau}$, $\sigma_1^2, \ldots, \sigma_R^2$, and $\boldsymbol{\Psi}$. If not specified, it uses the values in [51].
  Usage: `difm.hyp.parm(model.attributes, n.tau, n.s2.tau, n.sigma, n.s2.sigma, Hlist, Psi.size)`

- `difm.model.attributes`: Initializes the basic parameters and model attributes for DIFM. It requires information of the data, number of iterations, number of factors and the evolution matrix.
  Usage: `difm.model.attributes(data, n.iter, n.factors, G0)`

- `DIFMcpp`: Runs Dynamic ICAR Spatiotemporal Factors Model. The parameters are simulated from C++ codes. Users can control thin-in of iterations. Computation speed is faster than R.
  Usage: `DIFMcpp(model.attributes, hyp.parm, data, every = 1, verbose = TRUE)`

- `DIFMR`: Runs Dynamic ICAR Spatiotemporal Factors Model. The parameters are simulated from R codes. Users can control thin-in of iterations.
  Usage: `DIFMR(model.attributes, hyp.parm, data, every, verbose)`

- `marginal.d`: Calculates the Laplace-Metropolis predictive density [34] of the DIFM sample. Usage: `marginal.d(data, model.attributes, hyp.parm, Gibbs, burnin, verbose)`

- `marginal_d_cpp`: Calculates the Laplace-Metropolis predictive density for DIFM, using C++ approaches.
  Usage: `marginal_d_cpp(data, model.attributes, hyp.parm, Gibbs, burnin, verbose)`

- `permutation.order`: Sorts the variables according to the largest absolute value in the corresponding eigenvectors. If the variable was already selected previously from that turn, it selects the variable with next largest value.
  Usage: `permutation.order(data, n.factors)`

- `permutation.scale`: Permutes the dataset according to the absolute values of the eigenvectors and standardizes it. It sorts the variables as it is done in `permutation.order`.
  Usage: `permutation.scale(data, n.factors, return.scale)`

- `plot.B.CI`: Generates 95% credible intervals plot of factor loadings, $\mathbf{B}$.
  Usage: `plot.B.CI(Gibbs, true.val, burnin, permutation, main.bool, layout.dim)`

- `plot.B.spatial`: Outputs spatial plot of the factor loadings, $\mathbf{B}$. Red indicates positive values and blue indicates negative values.
  Usage: `plot.B.spatial(Gibbs, areapoly, burnin, permutation, main.bool, layout.dim)`

- `plot.sigma2.CI`: Generates 95% credible interval plot of the idiosyncratic variances, $\sigma_1^2, \ldots, \sigma_R^2$.
  Usage: `plot.sigma2.CI(Gibbs, burnin, permutation, main.bool)`

- `plot.tau.CI`: Returns 95% credible interval plot of the factor loadings variances, $\tau_1, \ldots, \tau_F$.
  Usage: `plot.tau.CI(Gibbs, burnin, true.val, main.bool)`

- `plot.X.CI`: Outputs 95% credible interval plot of the common factors, $\mathbf{X}$.
  Usage: `plot.X.CI(Gibbs, burnin, main.bool, layout.dim)`

In addition, the package includes two datasets for examples in the vignette. The datasets contain the number of violent crimes and property crimes per 100,000 people in western states in United States, respectively. The data was collected by Bureau of Justice Statistics annually from 1960 to 2019. Each dataset consists of 11 columns corresponding to the following states: Arizona, California, Colorado, Idaho, Montana, Nevada, Nevada, New Mexico, Oregon, Utah, Washington and Wyoming. In violent crime data, most of the states showed peak around the 1990s, followed by decrease or stagnation. In property crime data, majority of the states displayed peak in the 1980s, followed by sharp decrease from the 2000s.

4.2.2 Vignette of DIFM Package

We provide guidance for researchers through a vignette that provides an overview of the `DIFM` package. The first section gives a brief introduction of the `DIFM` package. In the second

section, we specify the model that incorporates dynamic linear model on the common factors, the ICAR prior on factor loadings matrix, and the hierarchical structural constraint. In the third section, we demonstrate two examples when applying DIFM to datasets of property crime rates and violent crime rates in the western states of United States.

DIFM:Dynamic ICAR Spatiotemporal Factor Models

Hwasoo Shin, Marco A. R. Ferreira

Introduction

This package `DIFM` provides codes to run Dynamic ICAR Spatiotemporal Factor Models (DIFM). It includes codes to initialize parameters, run MCMC, evaluate models, and generate plots. Read (Shin and Ferreira, 2023) for more details.

The vignette presents the model description of `DIFM` and the assumptions for common factors and factor loadings. We provide an example of crime rates in the western states of United States.

Model Description

We assume that the region of study is partitioned into r subregions. In our motivating example, the subregions are the states in the contiguous United States. The variable of interest in each subregion is observed at n time points. Let $\mathbf{y}_t$ be the r-dimensional vector of observations at time t ($t = 1, 2, \ldots, n$.)

We assume that the spatiotemporal behavior of the r subregions can be represented by k factors, where usually k is much smaller than r. Specifically, we assume the model

$$\mathbf{y}_t = \mathbf{B}\mathbf{x}_t + \mathbf{v}_t, \tag{1}$$

where $\mathbf{x}_t$ is the k-dimensional vector of factors at time t, $\mathbf{B}$ is an $r \times k$ matrix of factor loadings, and $\mathbf{v}_t$ is the r-dimensional vector of errors at time t. We assume that the observational error vector $\mathbf{v}_t$, $t = 1, 2, \ldots, n$ is independent over time and follows a Gaussian distribution $\mathbf{v}_t \sim \mathrm{N}(0, \mathbf{V})$, where $\mathbf{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_r^2)$. Each of the variances $\sigma_1^2, \ldots, \sigma_r^2$ is specific to one of the r subregions, and thus they are known as idiosyncratic variances.

We assume that the vector of factors $\mathbf{x}_t$ follows a dynamic linear model (West and Harrison, 1997; Prado et al., 2021). Specifically, we assume the general model

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}\theta_t, & (2)\\ \theta_t &= \mathbf{G}\theta_{t-1} + \omega_t, \omega_t \sim \mathrm{N}(0, \mathbf{W}), & (3) \end{aligned}$$

where θ_t is a latent process that allows great flexibility in the description of the temporal evolution of $\mathbf{x}_t$. Specifically, θ_t may encode different types of temporal trends as well as seasonality. For example, in our application we assume a second-order polynomial DLM and specify θ_t as a vector of dimension $2k$ that contains the level and the gradient of $\mathbf{x}_t$ at time t. In addition, the evolution matrix $\mathbf{G}$ describes the temporal evolution of the latent process θ_t. Further, ω_t is a $2k$-dimensional innovation vector with a dense covariance matrix $\mathbf{W}$. Finally, the matrix $\mathbf{F}$ relates the vector of common factors $\mathbf{x}_t$ to the appropriate elements of the latent process θ_t.

In the case of the second-order polynomial DLM that we consider, $\theta_t = (\theta_{t,1}, \theta_{t,2}, \ldots, \theta_{t,2k})^T$ is a vector of dimension $2k$ where $(\theta_{t,1}, \theta_{t,3}, \ldots, \theta_{t,2k-1})^T$ and $(\theta_{t,2}, \theta_{t,4}, \ldots, \theta_{t,2k})^T$ are respectively the level and the gradient of the vector of common factors $\mathbf{x}_t$. Thus, the matrix $\mathbf{F}$ that relates $\mathbf{x}_t$ to θ_t is a $k \times 2k$ matrix of

the form

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

The evolution matrix $\mathbf{G}$ has dimension $2k \times 2k$ and satisfies $\theta_{t,2j-1} = \theta_{t-1,2j-1} + \theta_{t-1,2j} + \omega_{t,2j-1}$ and $\theta_{t,2j} = \theta_{t-1,2j} + \omega_{t,2j}$, $j = 1, \dots, k$. Therefore, $\mathbf{G} = blockdiag(\mathbf{G}_0, \dots, \mathbf{G}_0)$ where

$$\mathbf{G}_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The specification of the factor loadings matrix $\mathbf{B}$ is crucial in our dynamic ICAR factor model. An important point to consider is the need for constraints on the matrix $\mathbf{B}$ to ensure identifiability of the model. Specifically, for any invertible $k \times k$ matrix $\mathbf{A}$, substituting $\mathbf{B}$ and $\mathbf{x}_t$ in Equation (1) by, respectively, $\mathbf{B}^* = \mathbf{BA}$ and $\mathbf{x}_t^* = \mathbf{A}^{-1}\mathbf{x}_t$ would lead to the same model. To ensure identifiability, we impose a hierarchical structural constraint that assumes that $\mathbf{B}$ is a full-rank block lower triangular matrix with diagonal elements equal to 1 (Aguilar and West, 2000). Specifically, we assume $\mathbf{B}$ has the form

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ b_{2,1} & 1 & 0 & \dots & 0 \\ b_{3,1} & b_{3,2} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{k,1} & b_{k,2} & b_{k,3} & \dots & 1 \\ b_{k+1,1} & b_{k+1,2} & b_{k+1,3} & \dots & b_{k+1,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{r,1} & b_{r,2} & b_{r,3} & \dots & b_{r,k} \end{bmatrix}.$$

To account for the spatial dependence among the factor loadings for neighboring subregions, we assume that each column of the matrix of factor loadings $\mathbf{B}$ follows an intrinsic conditional autoregressive model (Besag et al., 1991; Keefe et al., 2018). Specifically, we assume for the jth column $\mathbf{B}_j, j = 1, \dots, k$, the density

$$p(\mathbf{B}_j) \propto \exp\left(-\frac{1}{2\tau_j}\mathbf{B}_j^T\mathbf{H}\mathbf{B}_j\right), \tag{4}$$

where $\mathbf{H}$ is a precision matrix that accounts for the spatial dependence among neighboring subregions and τ_j controls the strength of spatial correlation among factor loadings. Specifically, if subregions i and j are neighbors, then the corresponding element of the matrix $\mathbf{H}$ is $h_{ij} = -g_{ij}$ where g_{ij} measures the strength of the association between subregions i and j. If subregions i and j are not neighbors, then $g_{ij} = 0$. Finally, the ith diagonal element of matrix $\mathbf{H}$ is $h_{ii} = \sum_{j \neq i} g_{ij}$. For example, a widely used choice for $\mathbf{H}$ assumes $g_{ij} = 1$ if i and j share a border, and $g_{ij} = 0$ otherwise. In that case, h_{ii} is equal to the number of neighbors of subregion i. Further, we assume that there are no islands which implies that the matrix $\mathbf{H}$ has one eigenvalue equal to 0 and all other eigenvalues larger than zero. Note that we assume this prior for each column of $\mathbf{B}$. Let $\mathbf{B}_{\cdot j}^* = \mathbf{B}_{(j+1):r,j}$ be the jth column of $\mathbf{B}$ without the first j elements that are fixed. In addition, let $\mathbf{H}_j^* = \mathbf{H}_{(j+1):r,(j+1):r}$. Then, the conditional distribution of $\mathbf{B}_{\cdot j}^*$ given $\mathbf{B}_{1:j,j} = (0, \dots, 0, 1)^T$ is multivariate normal with mean vector $\mathbf{h}_j = -\mathbf{H}_j^{*-1}\mathbf{H}_{(j+1):r,j}$ and precision matrix $\mathbf{H}_j^*$.

Examples

We apply DIFM to western United States crime datasets. The data was collected from Bureau of Justice Statistics and available at disaster center website. In this vignette, we provide two datasets, `Violent` and `Property` for violent and property crime, respectively. The data was collected from 50 states of United

States and District of Columbia from 1960 to 2019. In this example, we use the `WestStates` data included in `DIFM` package that contains the information of the map and polygon of the 11 western states: Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington and Wyoming. The numbers represent the cases of crime per 100,000 people. We use the square root of the data to stabilize the variance.

Step 1: Read and explore the data

```
data(Violent)
data(Property)
data(WestStates)
Violent <- as.matrix(Violent)
Violent <- sqrt(Violent)
Property <- as.matrix(Property)
Property <- sqrt(Property)
```

After we call the datasets, we explore the data through plots.

```
par(mar = c(2, 2, 2, 2))
layout(rbind(1:4, 5:8, c(9:11, 0)))
for (i in 1:11) {
    plot(1960:2019, Violent[, i], main = colnames(Violent)[i], type = "l", xlab = "",
        ylab = "")
}
```

Figure 1 shows the square root of the number of violent crimes in Western states. In most of the states, the cases of violent crimes soar by 2000 and have small changes. The trend after 2000 differ by states. Since many states share similar trend, we can assume that DIFM would be applicable in this case.

```
par(mar = c(2, 2, 2, 2))
layout(rbind(1:4, 5:8, c(9:11, 0)))
for (i in 1:11) {
    plot(1960:2019, Property[, i], main = colnames(Property)[i], type = "l", xlab = "",
        ylab = "")
}
```

Figure 2 shows the square root of the number of property crimes in Western states. Property crimes soar by 1980 in western states, but would usually decrease from then. Many states show start of sharp decrease between 1980 and 1990. These information can be reperesented with smaller number of factors through DIFM.

Step 2: Run DIFM with range of factors.

Now we run DIFM with different number of factors. For our two examples, we try models from 1 to 4 factors. Before we start DIFM, we should permute the order of the variable to adjust the structural hierarchical constraint. We set the variable that would represent the factor well according to the eigenvectors. From Figure 1 and Figure 2, we can find that the time serires are rather non-stationary than stationary. Therefore, we consider a second order polynomial for the dynamic linear model. First, we run MCMC for the violent crime data.

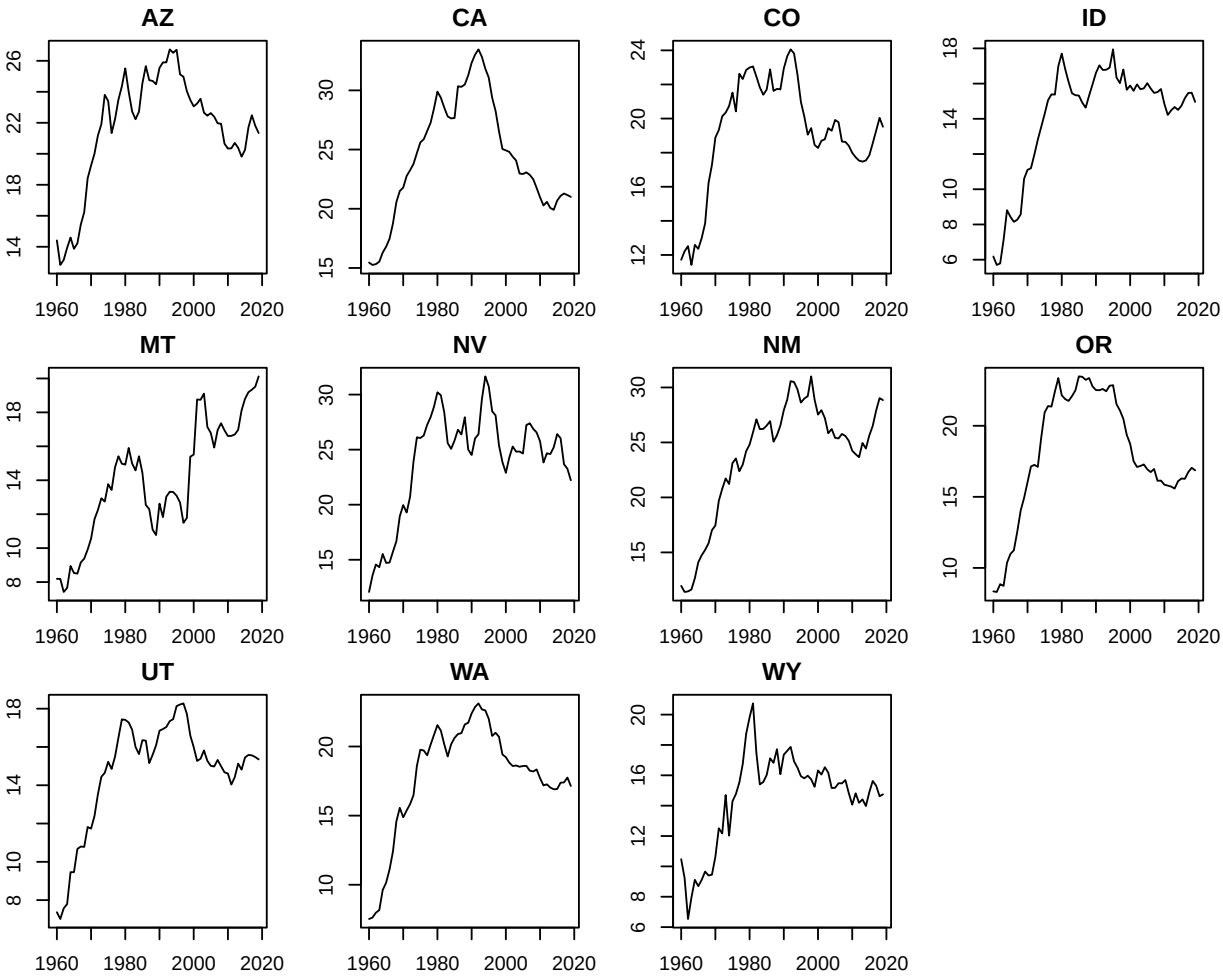


Figure 1: Figure 1: Number of violent crimes

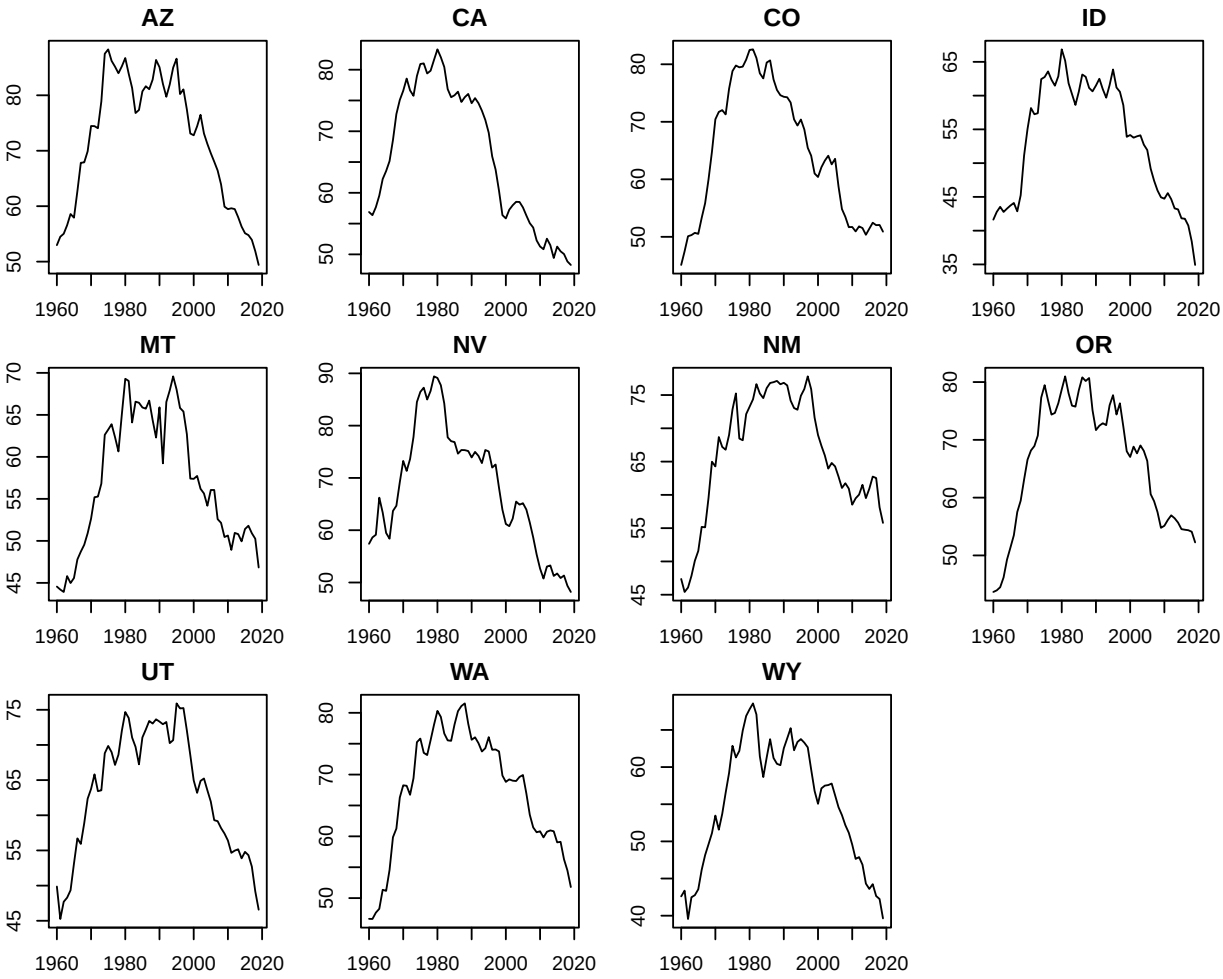


Figure 2: Figure 2: Number of property crimes

```
n.iter <- 5000
n.save <- 10
G0 <- rbind(c(1, 1), c(0, 1))
Violent.permutation <- permutation.order(Violent, 4)
Violent <- Violent[, Violent.permutation]
Violent.Hlist <- buildH(WestStates, Violent.permutation)

set.seed(1101)

model.attributes1V <- difm.model.attributes(Violent, n.iter, n.factors = 1, G0)
hyp.parm1V <- difm.hyp.parm(model.attributes1V, Hlist = Violent.Hlist)
ViolentDIFM1 <- DIFMcpp(model.attributes1V, hyp.parm1V, Violent, every = n.save,
    verbose = FALSE)
ViolentAssess1 <- marginal_d_cpp(Violent, model.attributes1V, hyp.parm1V, ViolentDIFM1,
    verbose = FALSE)

model.attributes2V <- difm.model.attributes(Violent, n.iter, n.factors = 2, G0)
hyp.parm2V <- difm.hyp.parm(model.attributes2V, Hlist = Violent.Hlist)
ViolentDIFM2 <- DIFMcpp(model.attributes2V, hyp.parm2V, Violent, every = n.save,
    verbose = FALSE)
ViolentAssess2 <- marginal_d_cpp(Violent, model.attributes2V, hyp.parm2V, ViolentDIFM2,
    verbose = FALSE)

model.attributes3V <- difm.model.attributes(Violent, n.iter, n.factors = 3, G0)
hyp.parm3V <- difm.hyp.parm(model.attributes3V, Hlist = Violent.Hlist)
ViolentDIFM3 <- DIFMcpp(model.attributes3V, hyp.parm3V, Violent, every = n.save,
    verbose = FALSE)
ViolentAssess3 <- marginal_d_cpp(Violent, model.attributes3V, hyp.parm3V, ViolentDIFM3,
    verbose = FALSE)

model.attributes4V <- difm.model.attributes(Violent, n.iter, n.factors = 4, G0)
hyp.parm4V <- difm.hyp.parm(model.attributes4V, Hlist = Violent.Hlist)
ViolentDIFM4 <- DIFMcpp(model.attributes4V, hyp.parm4V, Violent, every = n.save,
    verbose = FALSE)
ViolentAssess4 <- marginal.d(Violent, model.attributes4V, hyp.parm4V, ViolentDIFM4,
    verbose = FALSE)
```

Now we run the MCMC for the property crime data.

```
Property.permutation <- permutation.order(Property, 4)
Property <- Property[, Property.permutation]
Property.Hlist <- buildH(WestStates, Property.permutation)

set.seed(1101)

model.attributes1P <- difm.model.attributes(Property, n.iter, n.factors = 1, G0)
hyp.parm1P <- difm.hyp.parm(model.attributes1P, Hlist = Property.Hlist)
PropertyDIFM1 <- DIFMcpp(model.attributes1P, hyp.parm1P, Property, every = n.save,
    verbose = FALSE)
PropertyAssess1 <- marginal_d_cpp(Property, model.attributes1P, hyp.parm1P, PropertyDIFM1,
    verbose = FALSE)

model.attributes2P <- difm.model.attributes(Property, n.iter, n.factors = 2, G0)
```

```
hyp.parm2P <- difm.hyp.parm(model.attributes2P, Hlist = Property.Hlist)
PropertyDIFM2 <- DIFMcpp(model.attributes2P, hyp.parm2P, Property, every = n.save,
    verbose = FALSE)
PropertyAssess2 <- marginal_d_cpp(Property, model.attributes2P, hyp.parm2P, PropertyDIFM2,
    verbose = FALSE)

model.attributes3P <- difm.model.attributes(Property, n.iter, n.factors = 3, G0)
hyp.parm3P <- difm.hyp.parm(model.attributes3P, Hlist = Property.Hlist)
PropertyDIFM3 <- DIFMcpp(model.attributes3P, hyp.parm3P, Property, every = n.save,
    verbose = FALSE)
PropertyAssess3 <- marginal_d_cpp(Property, model.attributes3P, hyp.parm3P, PropertyDIFM3,
    verbose = FALSE)

model.attributes4P <- difm.model.attributes(Property, n.iter, n.factors = 4, G0)
hyp.parm4P <- difm.hyp.parm(model.attributes4P, Hlist = Property.Hlist)
PropertyDIFM4 <- DIFMcpp(model.attributes4P, hyp.parm4P, Property, every = n.save,
    verbose = FALSE)
PropertyAssess4 <- marginal.d(Property, model.attributes4P, hyp.parm4P, PropertyDIFM4,
    verbose = FALSE)
```

We select the best model through the Metropolis-Laplace estimator of the predictive density.

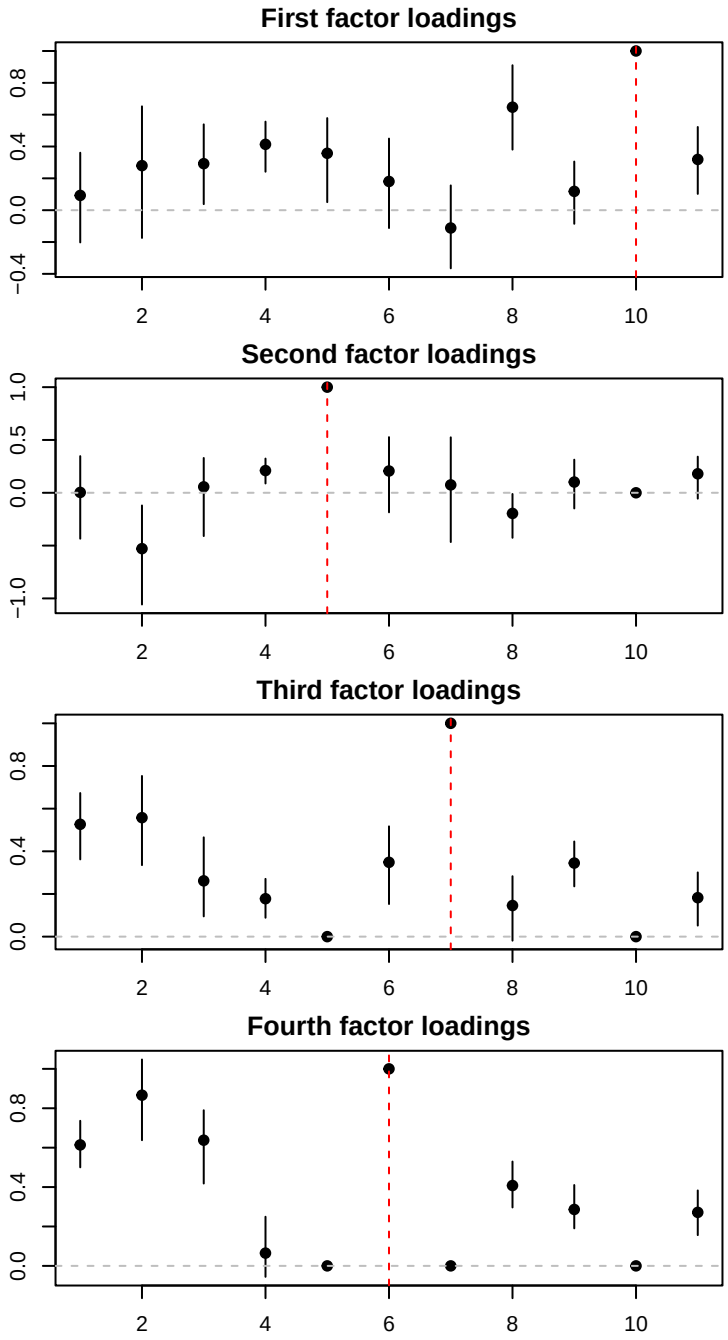
```
PDtable <- matrix(NA, 2, 4)
PDtable[1, ] <- c(ViolentAssess1$Maximum, ViolentAssess2$Maximum, ViolentAssess3$Maximum,
    ViolentAssess4$Maximum)
PDtable[2, ] <- c(PropertyAssess1$Maximum, PropertyAssess2$Maximum, PropertyAssess3$Maximum,
    PropertyAssess4$Maximum)
PDtable <- as.data.frame(PDtable)
rownames(PDtable) <- c("Violent", "Property")
colnames(PDtable) <- paste("Factors =", 1:4)
kable(PDtable)
```

	Factors = 1	Factors = 2	Factors = 3	Factors = 4
Violent	-1422.399	-1408.117	-1415.552	-1548.975
Property	-2083.176	-2135.292	-2200.640	-2420.407

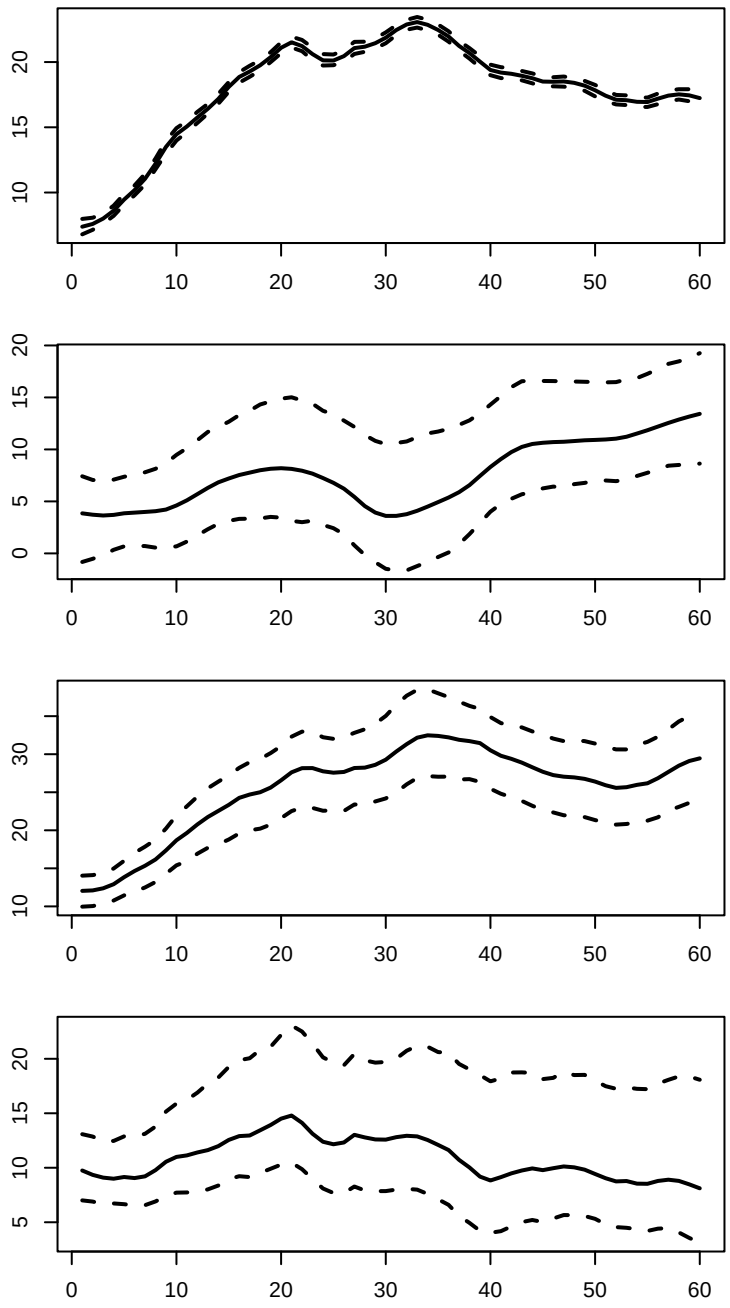
From the table, we can find that the best models according to the evaluation method are 2-factor-model for violent crime and 1-factor-model for property crime datasets.

We present a few posterior distribution plots of the 4-factor models. The plots below are from the violent crime data.

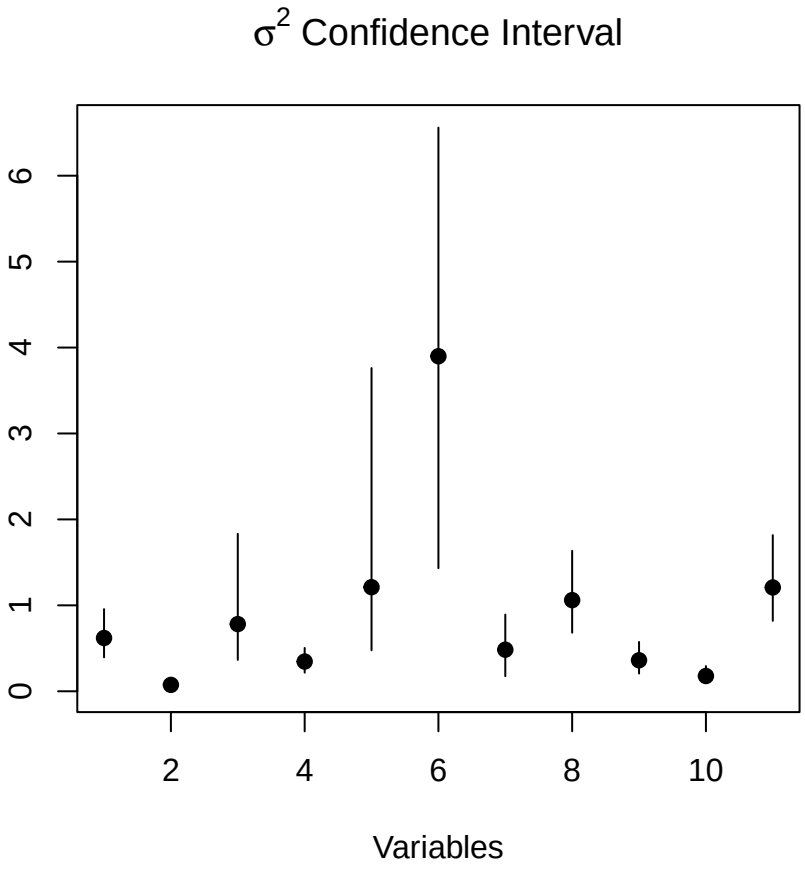
```
par(mar = c(2, 2, 2, 2))
plot.B.CI(ViolentDIFM4, permutation = Violent.permutation)
```



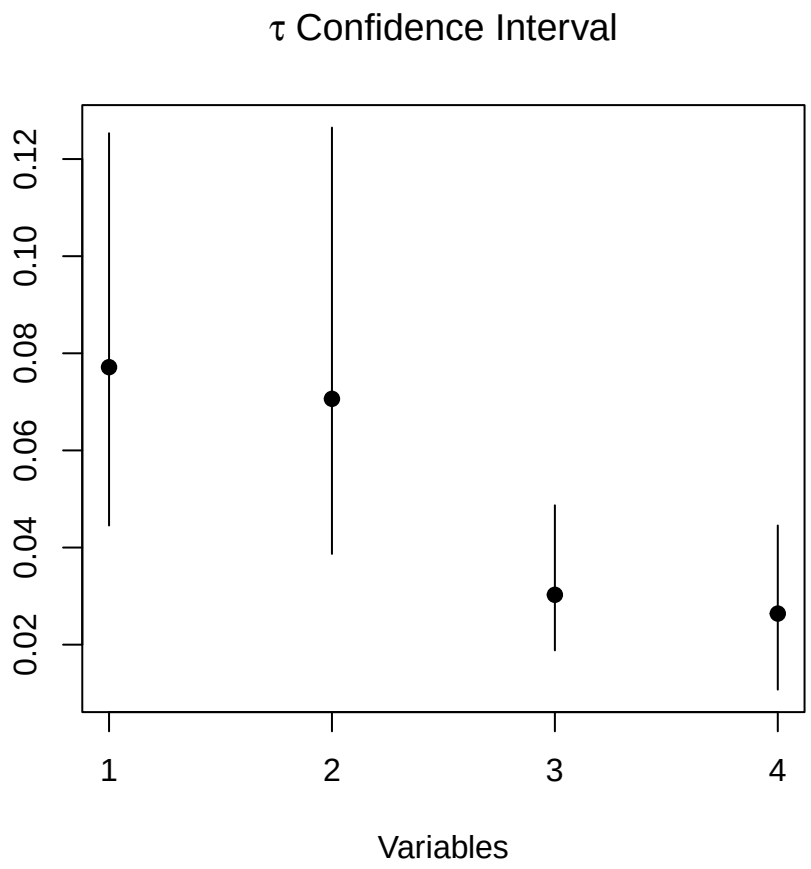

```
plot.X.CI(ViolentDIFM4)
```



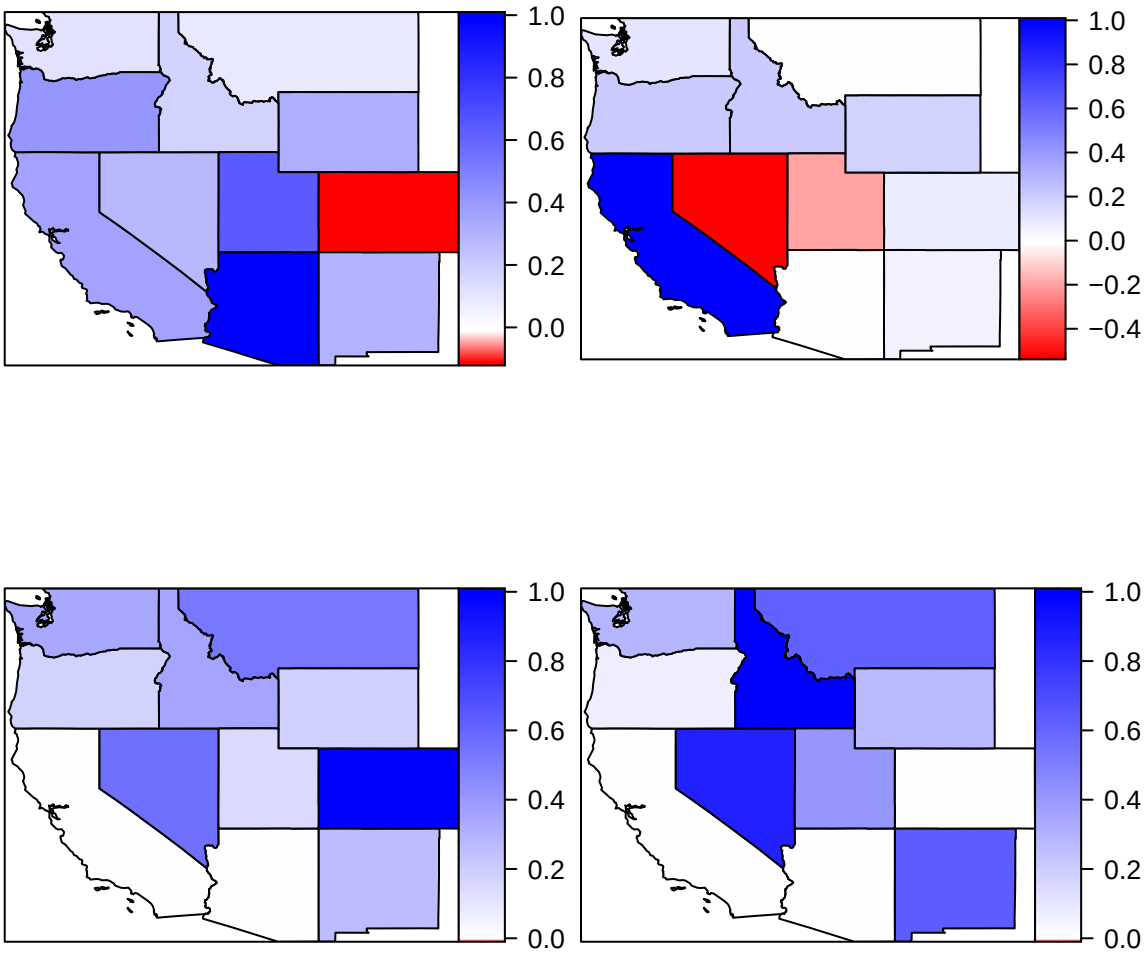

```
plot.sigma2.CI(ViolentDIFM4, permutation = Violent.permutation)
```




```
plot.tau.CI(ViolentDIFM4)
```

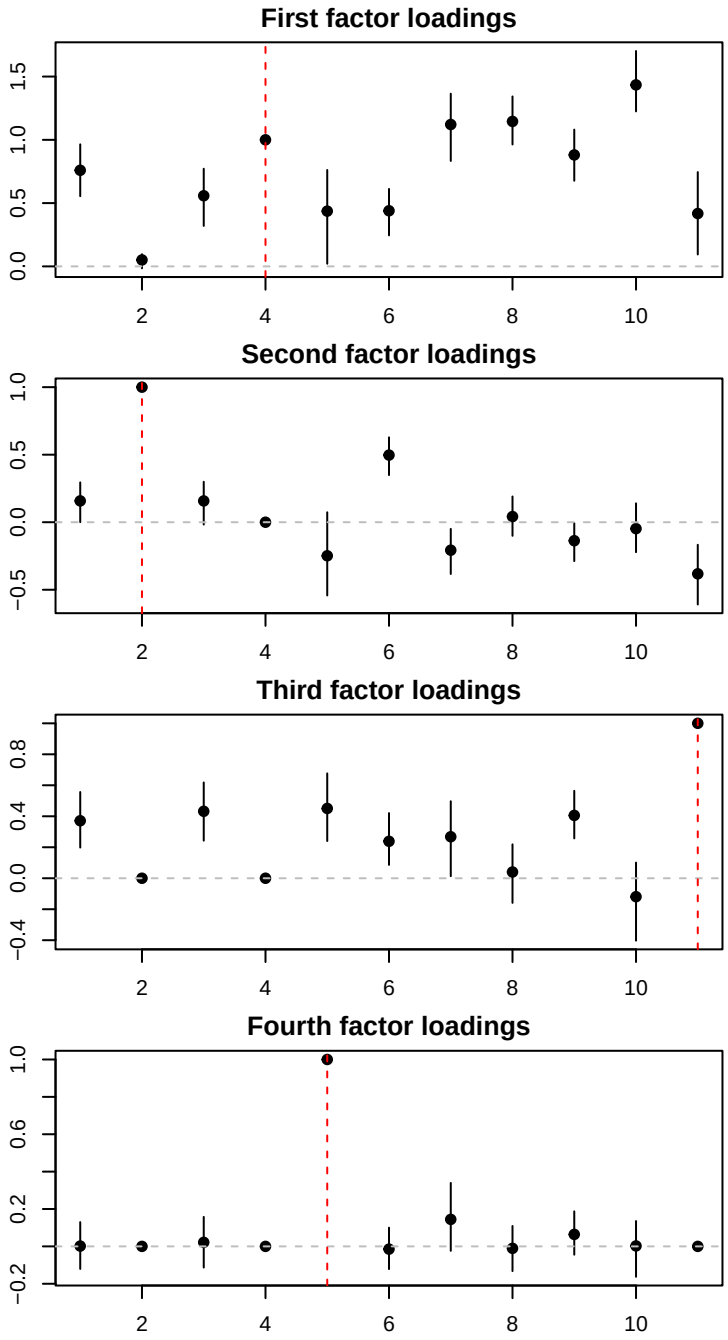



```
plot.B.spatial(ViolentDIFM4, WestStates, layout.dim = c(2, 2))
```

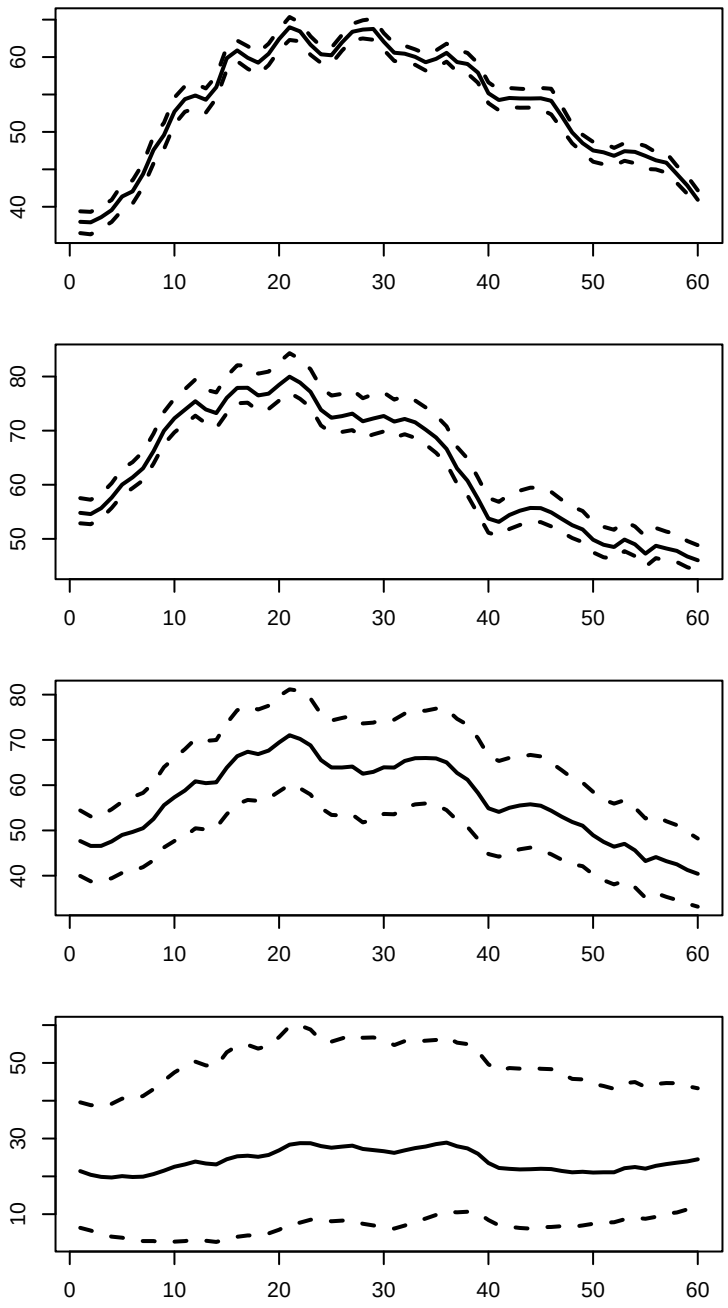


In the same manner, we present the results of the property crime data.

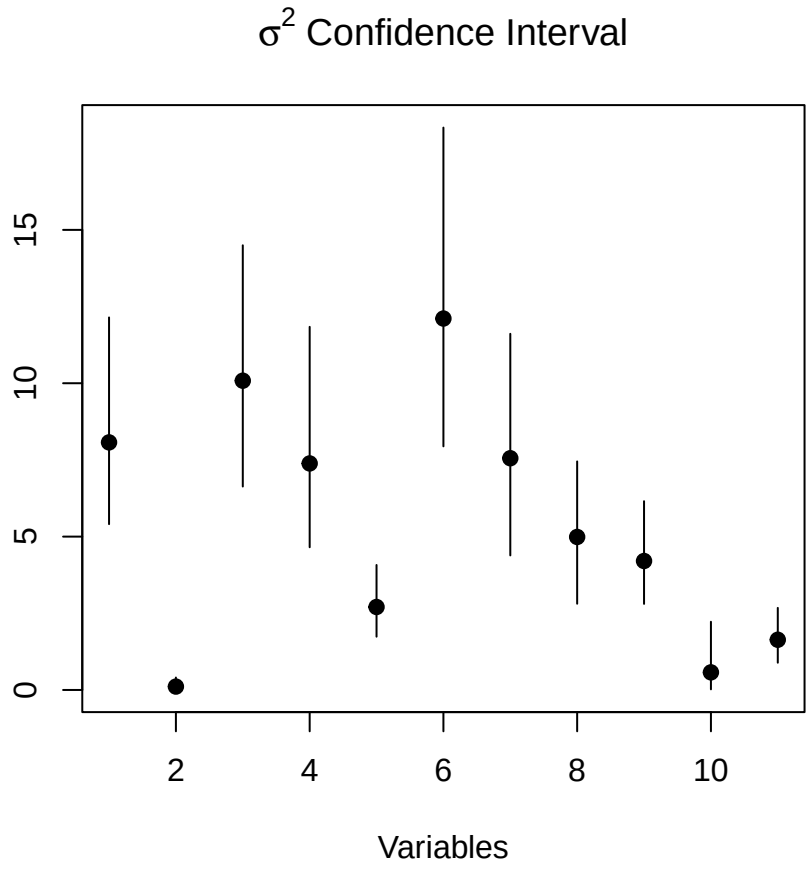
```
par(mar = c(2, 2, 2, 2))
plot.B.CI(PropertyDIFM4, permutation = Property.permutation)
```



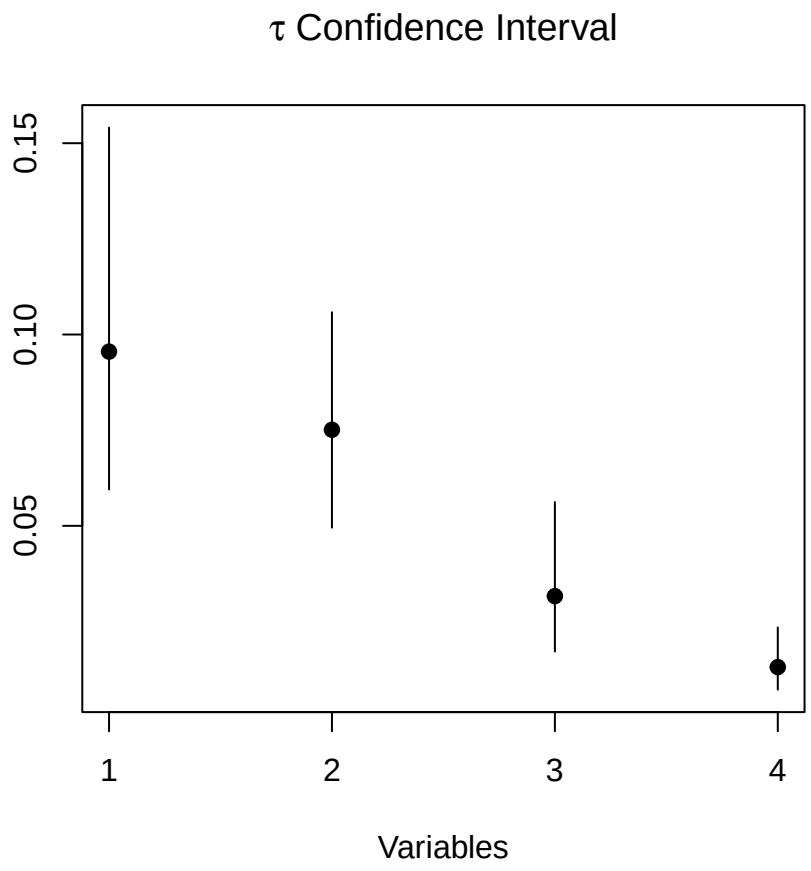

```
plot.X.CI(PropertyDIFM4)
```



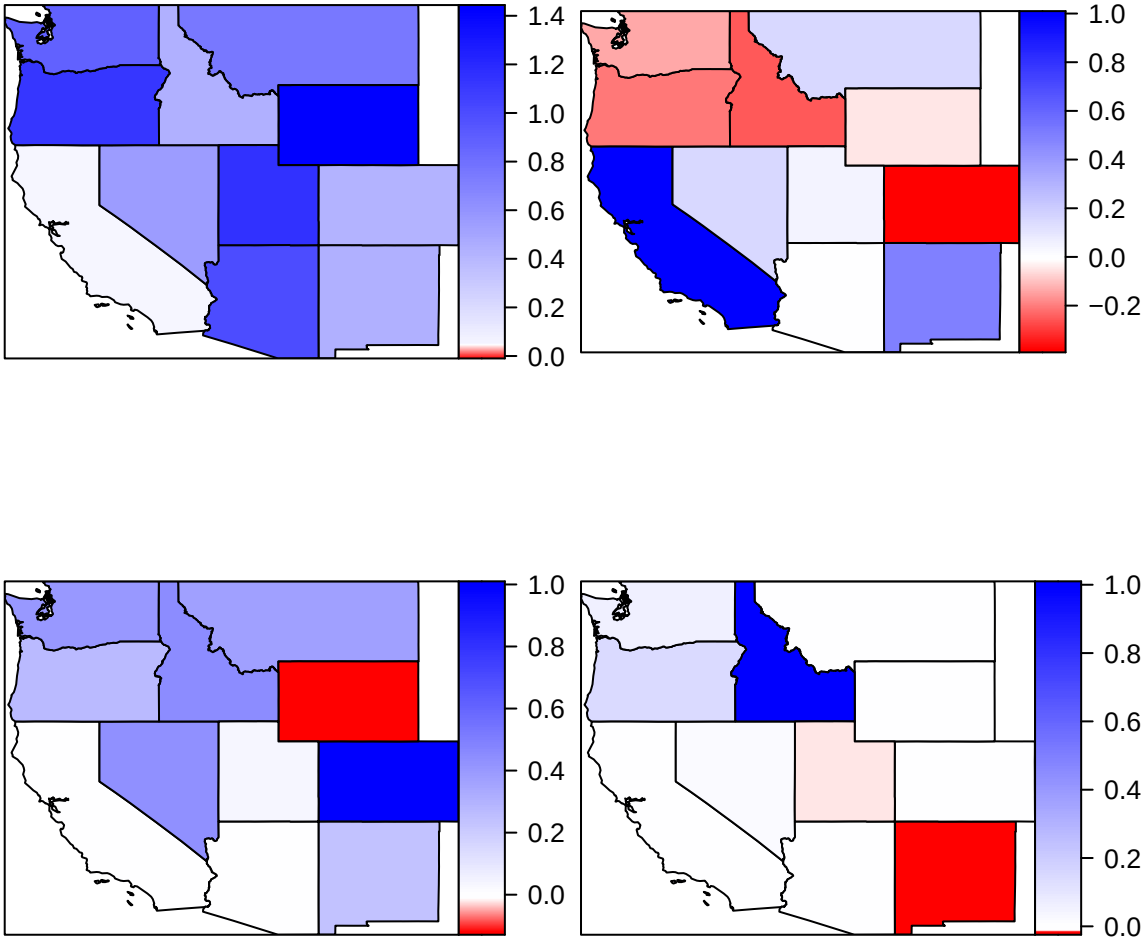

```
plot.sigma2.CI(PropertyDIFM4, permutation = Property.permutation)
```




```
plot.tau.CI(PropertyDIFM4)
```




```
plot.B.spatial(PropertyDIFM4, WestStates, layout.dim = c(2, 2))
```



Reference

- Shin, H. and Ferreira, M. A. (2023). "Dynamic ICAR Spatiotemporal Factor Models." Spatial Statistics, 56, 100763.
- West, M. and Harrison, J. (1997). Bayesian Forecasting and Dynamic Models (2nd Ed.). Berlin, Heidelberg: Springer-Verlag.
- Prado, R., Ferreira, M. A. R., and West, M. (2021). Time Series: Modeling, Computation, and Inference. Boca Raton (2nd Ed.): Chapman & Hall/CRC.
- Aguilar, O. and West, M. (2000). "Bayesian dynamic factor models and portfolio allocation." Journal of Business and Economic Statistics, 18, 338–357.
- Besag, J., York, J., and Mollie, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." Annals of the Institute of Statistical Mathematics, 43, 1
- Keefe, M. J., Ferreira, M. A. R., and Franck, C. T. (2018). "On the formal specification of sum-zero constrained intrinsic conditional autoregressive models." Spatial Statistics, 24.

4.3 Package: BCFM

The `BCFM` package offers functions for setting weakly informative hyperparameters, simulating MCMC samples, generating figures, and selecting the most appropriate models for the number of clusters and the number of factors. Parameters are sampled from their full conditionals as described in Chapter 3.

This package have several dependencies: `Rcpp, RcppArmadillo, ggplot2, LaplacesDemon, ggsci, ggpubr, fastmatrix` and `tidyr`. We call C++ algorithms and functions through `Rcpp` [17] and `RcppArmadillo` [18] packages. All plots are generated with `ggplot2` package [59]. We simulate random parameters that follows inverse gamma and inverse Wishart distribution from `LaplacesDemon` [52]. The color scheme of this package follows the guideline of *Journal of Oncology*, and we use the package `ggsci` to apply the colors. We use `ggpubr` to extract the legends of the plots and arrange them with user-specified layout. For LDL decomposition to derive the hyperparametrs for cluster means and covariances, we use `fastmatrix` [42] package. Finally, we utilize functions in `tidyr` to organize the dataset.

4.3.1 Documentation of BCFM Package

This package contains 43 R functions and 9 C++ functions. Among these codes, there are 26 R functions accessible to users. The functions that are available for users are in the following list.

- `BCFMcpp`: Runs the MCMC of BCFM using the C++ functions.
  Usage: `BCFMcpp(data, model.attributes, hyp.parm, n.iter, every, verbose)`
- `BCFMR`: Runs the MCMC of BCFM using R functions.
  Usage: `BCFMR(data, model.attributes, hyp.parm, n.iter, every, verbose)`
- `BIC.like`: Computes the BIC with integrated likelihood criterion of the model. The model demonstrates good performance with lower values of this criterion.
  Usage: `BIC.like(data, Gibbs, model.attributes, burnin)`
- `getmode`: Computes the mode of the given vector.
  Usage: `getmode(v)`

- `ggplot.B.CI`: Outputs 95% credible intervals of the factor loadings matrix, $\mathbf{B}$. Each plot corresponds to its factor. Note that fixed variable at the factor will not have credible intervals.
  Usage: `ggplot.B.CI(Gibbs, true.val, burnin, permutation, layout.dim)`

- `ggplot.B.trace`: Generates a trace plot of the factor loadings from one variable. It shows the sampled parameters of the factor loadings of a variable specified by the user.
  Usage: `ggplot.B.trace(Gibbs, burnin, permutation, true.val, factor.num)`

- `ggplot.mu.density`: Draws a posterior density plot of the cluster means, $\boldsymbol{\mu}$. Each plot represents the factor and colors represent clusters.
  Usage: `ggplot.mu.density(Gibbs, true.val, add.legend, burnin, layout.dim)`

- `ggplot.mu.trace`: Returns a trace plot of the cluster means of the factor. The function outputs a plot of the simulated cluster means of a factor set by the user.
  Usage: `ggplot.mu.trace(Gibbs, true.val, burnin, factor.num)`

- `ggplot.Omega.density`: Generates a posterior density plot of the cluster covariance, $\boldsymbol{\Omega}$. It returns the diagonal and the off-diagonal values of a covariance matrix from the selected cluster.
  Usage: `ggplot.Omega.density(Gibbs, group.num, true.val, burnin)`

- `ggplot.Omega.trace`: Outputs a trace plot of the cluster covariance from the selected group.
  Usage: `ggplot.Omega.trace(Gibbs, burnin, group.num)`

- `ggplot.probs.density`: Yields posterior density plot of the cluster assignment probabilities, $\mathbf{p}$.
  Usage: `ggplot.probs.density(Gibbs, burnin, truep)`

- `ggplot.probs.trace`: Draws a trace plot of the simulated values of the cluster probabilities.
  Usage: `ggplot.probs.trace(Gibbs, burnin)`

- `ggplot.sigma2.CI`: Generates a 95% credible intervals plot of the idiosyncratic variances, $\sigma_1^2, \ldots, \sigma_R^2$.
  Usage: `ggplot.sigma2.CI(Gibbs, burnin, permutation, true.val)`

- `ggplot.sigma2.trace`: Returns a trace plot of the idiosyncratic variance from the specific variable researcher sets.
  Usage: `ggplot.sigma2.trace(Gibbs, burnin, permutation, variable.num)`

- `ggplot.tau.CI`: Outputs a trace plot of the factor loadings variance, $\tau_1, \ldots, \tau_F$.
  Usage: `ggplot.tau.CI(Gibbs, burnin, true.val)`

- `ggplot.tau.trace`: Illustrates a trace plot of the factor loadings variance.
  Usage: `ggplot.tau.trace(Gibbs, burnin, true.val)`

- `ggplot.Zit.heatmap`: Yields a heatmap of the cluster assignments, $\mathbf{Z}$. The observations are plotted according to the largest cluster assignment probability, and their true clusters when available.
  Usage: `ggplot.Zit.heatmap(Gibbs, true.val, burnin)`

- `initialize.hyp.parm`: Initializes the hyperparameters for BCFM. It sets hyperparameters of the idiosyncratic variances, factor loadings variance, cluster means, cluster covariances, and cluster assignment probabilities. If not specified by the user, it uses the information in the Bayesian Clustering Factor Models paper.
  Usage: `initialize.hyp.parm(data, model.attributes, n.sigma, n.s2.sigma, n.tau,`
  `n.s2.tau, omega.diag.nu, p.exponent, covariance, diag.Psi, vague.mu, zero.mu)`

- `initialize.model.attributes`: Sets the model attributes of BCFM. It requires the following information: number of observations, number of variables, number of timepoints, number of clusters, and number of factors.
  Usage: `initialize.model.attribute(S, times, R, L, G)`

- `initialize.model.parameters`: Initializes the model parameters for a simulated dataset. It requires users to set the model attributes from `initialize.model.attributes` prior to this function.
  Usage: `initialize.model.parameters(model.attributes, c.probs, model.means, model.omega, model.sigmaV, model.taus, model.B`

- `marginal.d.mean.transformed`: Computes the Laplace-Metropolis estimator of the marginal density [34]. To consider the approximation of Gaussian distribution, we

transform cluster probabilities into logit values, and idiosyncratic variances and factor loadings covariance into log values. We use the posterior mean of the simulated parameters to compute this criterion.
Usage: `marginal.d.mean.transformed(data.row, Gibbs, model.attributes, hyp.parm, burnin)`

- `permutation.scale`: Returns the permutation order and the data according to the largest absolute values of eigenvectors. This function is necessary to sort the variables prior to running MCMC. It also includes an option to standardize the dataset.
  Usage: `permutation.scale(data, permutation, covariance, return.array, num.layers, permutation.out`

- `simulated.data`: Simulates a dataset according to the model attributes from `initialize.model.attributes` and parameters set from `initialize.model.parameters`.
  Usage: `simulated.data(model.attributes, model.parm)`

- `sub.model.parameters`: Extracts the parameter information from the dataset and model attributes from `initialize.model.attributes`.
  Usage: `sub.model.attributes(data, model.attributes)`

- `swap.label`: Swaps the labels of a vector.
  Usage: `swap.label(label.vector, before, after)`

4.3.2 Vignette of BCFM Package

The `BCFM` package includes a vignette to give users insight how to use Bayesian clustering Factors Model. First, we introduce BCFM and how it can be applied to datasets. The second section outlines the model description of BCFM, which combines Bayesian factor models with Gaussian mixture models for clustering and hierarchical structural constraints for identifiability. Finally, in the third section, we present an example utilizing a simulated dataset with 4 clusters and 3 factors, demonstrating the usage of BCFM functions for running MCMC, generating plots, and evaluating the model.

BCFM Vignettes

Hwasoo Shin, Marco A. R. Ferreira, Allison N. Tegge

Introduction

This package provides R and C++ codes to run MCMC computations for Bayesian Clustering Factor Models (BCFM). BCFM is a combination of Bayesian factor models with clustering method, assuming the vector of observations can be written as a linear function of latent common factors. Usually, the number of the factors is much smaller than the dimension of the vector of observations. We apply Gaussian mixture model to find latent clusters in lower dimensional spaces.

Model Description

This section describes the model we propose for performing this concomitant dimension reduction and clustering. We consider a setting with multivariate R variables observed for each of n subjects. In addition, we assume that we can explain the dependence structure among the R variables with a much smaller number F of latent variables or factors. Further, we assume that in this smaller F-dimensional latent space, the subjects may be clustered into K clusters.

Let $\mathbf{y}_i$ be an R-dimensional vector with the R observed variables from subject i. We assume the following factor model with F factors

$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i + \mathbf{v}_i, \tag{1}$$

where $\mathbf{B}$ is an $R \times F$ matrix of factor loadings, and $\mathbf{x}_i$ is an F-dimensional vector of common factor for subject i. The error vector is $\mathbf{v}_i \sim \mathbf{N}(\mathbf{0}, \mathbf{V})$ with $\mathbf{V} = diag(\sigma_1^2, \ldots, \sigma_R^2)$ and $\sigma_1^2, \ldots, \sigma_R^2$ is the idiosyncratic variance. Note that Equation (1) encodes a dimension reduction from dimension R to dimension F. To ensure the identifiability of the model, we assume that the matrix of factor loadings $\mathbf{B}$ follows a hierarchical structural constraint (Prado et al., 2021; Aguilar and West, 2000). Specifically, $\mathbf{B}$ has the form

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ b_{2,1} & 1 & 0 & \ldots & 0 \\ b_{3,1} & b_{3,2} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{F,1} & b_{F,2} & b_{F,3} & \ldots & 1 \\ b_{F+1,1} & b_{F+1,2} & b_{F+1,3} & \ldots & b_{F+1,F} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{R,1} & b_{R,2} & b_{R,3} & \ldots & b_{R,F} \end{bmatrix}.$$

Thus, the matrix of factor loadings $\mathbf{B}$ is a lower triangular matrix with main diagonal elements are all equal to 1. Each row of $\mathbf{B}$ corresponds to an observed variable, and each column corresponds to a common factor. The order of the variables is crucial because of the properties of the hierarchical structural constraints.

We assume the common factors $\mathbf{x}_i$ follow a Gaussian mixture model. This allows the ith subject to be allocated to one of the K clusters. Let z_i indicate the cluster subject i belongs to. Then, given $z_i = k$, the common factor x_i has the Gaussian conditional distribution

$$\mathbf{x}_i | z_i = k \sim N(\mu_k, \mathbf{\Omega}_k), \tag{2}$$

where μ_k is the mean vector and $\mathbf{\Omega}_k$ is the covariance matrix of the common factors of the cluster k. Let the probability of a randomly selected subject assigned to cluster k be $P(z_i = k) = p_k$. Then, the Gaussian mixture model for $\mathbf{x}_i$ is

$$\mathbf{x}_i \sim \sum_{k=1}^{K} p_k N(\mu_k, \mathbf{\Omega}_k). \tag{3}$$

Note that to make the model to be identifiable, we impose another constraint that the largest cluster, which is the first cluster, has a diagonal covariance matrix.

Functions

In this section, we introduce the functions in this package.

- **initialize.model.attributes**: Initializes the model attributes used for the dataset and MCMC iterations. The user should specify the number of subjects (S), number of timepoints (times), number of factors (L), number of groups (G), and number of covariates (R).
- **initialize.hyp.parm**: Initializes the hyperparameters related to factors model and clustering.
- **initialize.model.parameters**: Initializes the model parameters when you don't have the actual data. Once you set the model attributes through `initialize.model.attributes`, it will simulate random dataset with parameters, unless you specify the true parameters.
- **simulated.data**: Simulates data based on the model attributes from `initialize.model.attributes` and true model parameters from `initialize.model.parameters`.
- **sub.model.parameters**: Extracts the model parameters information from a given dataset and model attributes from `initialize.model.attributes`.
- **permutation.scale**: Finds and outputs the permutation of the 3-dimensional data according to the eigen vectors. This function also provides the standardized the data.
- **BCFMR**: Runs MCMC of BCFM using R codes.
- **BCFMcpp**: Runs MCMC of BCFM using C++ codes.
- **BIC.like**: Computes the BIC-like criterion of the result from `BCFMR` or `BCFMcpp`. It uses the posterior mean of the simulated parameters.
- **marginal.d.mean.transformed**: Computes the Laplace-Metropolis likelihood of marginal density (Prado et al., 2021) of the result from `BCFMR` or `BCFMcpp`. It uses the posterior mean of the simulated parameters.
- **swap.label**: Swap the labels of the group.
- **ggplot.-.density**: A ggplot that simulates the density of the posteriors. Probabilities $(p_1, \ldots, p_G)$, cluster means (μ) and covariances $(\mathbf{\Omega})$ can be used.
- **ggplot.-.CI**: A ggplot that returns the credible interval plot of the posteriors. Factor loadings $(\mathbf{B})$, observational errors $(\sigma_1^2, \ldots, \sigma_R^2)$ and factor loadings variances $(\tau_1, \ldots, \tau_F)$ can be used.
- **ggplot.heatmap.Zit**: A ggplot that yields the heatmap of cluster assignments $(\mathbf{Z})$.
- **ggplot.-.trace**: A ggplot that draws the MCMC trace plots of the simulated parameters.

Example

We simulate an example where the data has 800 subjects and 10 variables. The true setting of the dataset has 4 clusters and 3 factors. The cluster assignment probabilities are (0.4, 0.3, 0.2, 0.1), and the mean of the each 4 cluster is (-6, -6, -6), (-2, -2, -2), (2, 2, 2), and (5, 5, 5). The covariance of the clusters are

$$\mathbf{\Omega}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{\Omega}_2 = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}, \mathbf{\Omega}_3 = \begin{bmatrix} 2 & -0.5 & -0.5 \\ -0.5 & 2 & -0.5 \\ -0.5 & -0.5 & 2 \end{bmatrix}, \mathbf{\Omega}_4 = \begin{bmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}.$$

The idiosyncratic variance of the variables are all equal to 0.1, and the variance of the factor loadings are all equal to 0.1. We first use the three functions `initialize.model.attributes`, `initialize.model.parameters` and `simulated.data` to simulate a sample data.

```
model.attributes <- initialize.model.attributes(S = 800, times = 1, R = 10, L = 3,
    G = 4)
c.probs <- c(0.4, 0.3, 0.2, 0.1)
model.means <- rbind(c(-5, -6, -6), c(-2, -3, -2), c(2, 2, 2), c(5, 5, 5))
model.omega <- array(NA, dim = c(4, 3, 3))
model.omega[1, , ] <- rbind(c(1, 0, 0), c(0, 1, 0), c(0, 0, 1))
model.omega[2, , ] <- rbind(c(1, 0.5, 0.5), c(0.5, 1, 0.5), c(0.5, 0.5, 1))
model.omega[3, , ] <- rbind(c(2, -0.5, -0.5), c(-0.5, 2, -0.5), c(-0.5, -0.5, 2))
model.omega[4, , ] <- rbind(c(1, -0.5, 0.5), c(-0.5, 1, -0.5), c(0.5, -0.5, 1))
model.sigmaV <- rep(0.1, 10)
model.taus <- rep(0.1, 3)

set.seed(110192)
model.parm <- initialize.model.parameters(model.attributes, c.probs, model.means,
    model.omega, model.sigmaV, model.taus, model.B = NA)
sim.data <- simulated.data(model.attributes, model.parm)
data <- sim.data$data
probs.true <- as.vector(table(sim.data$pselected)/nrow(sim.data$data))
```

Now we permute the dataset according to the eigenvectors and set the hyperparmaeters for the priors.

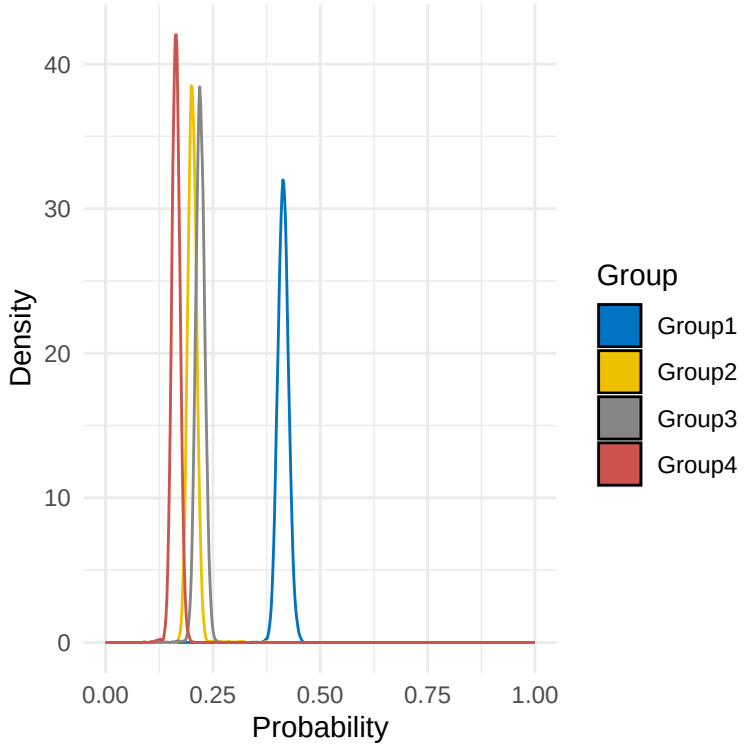
```
hyp.parm <- initialize.hyp.parm(data, model.attributes, p.exponent = c(2, 2, 2, 2),
    vague.mu = FALSE)
```

We run the MCMC with the CPP codes through `BCFMcpp`.

```
n.iter <- 10000
tic <- Sys.time()
simBCFM <- BCFMcpp(sim.data$data, model.attributes, hyp.parm, n.iter, every = 1,
    verbose = FALSE)
toc <- Sys.time()
run.time <- toc - tic
```

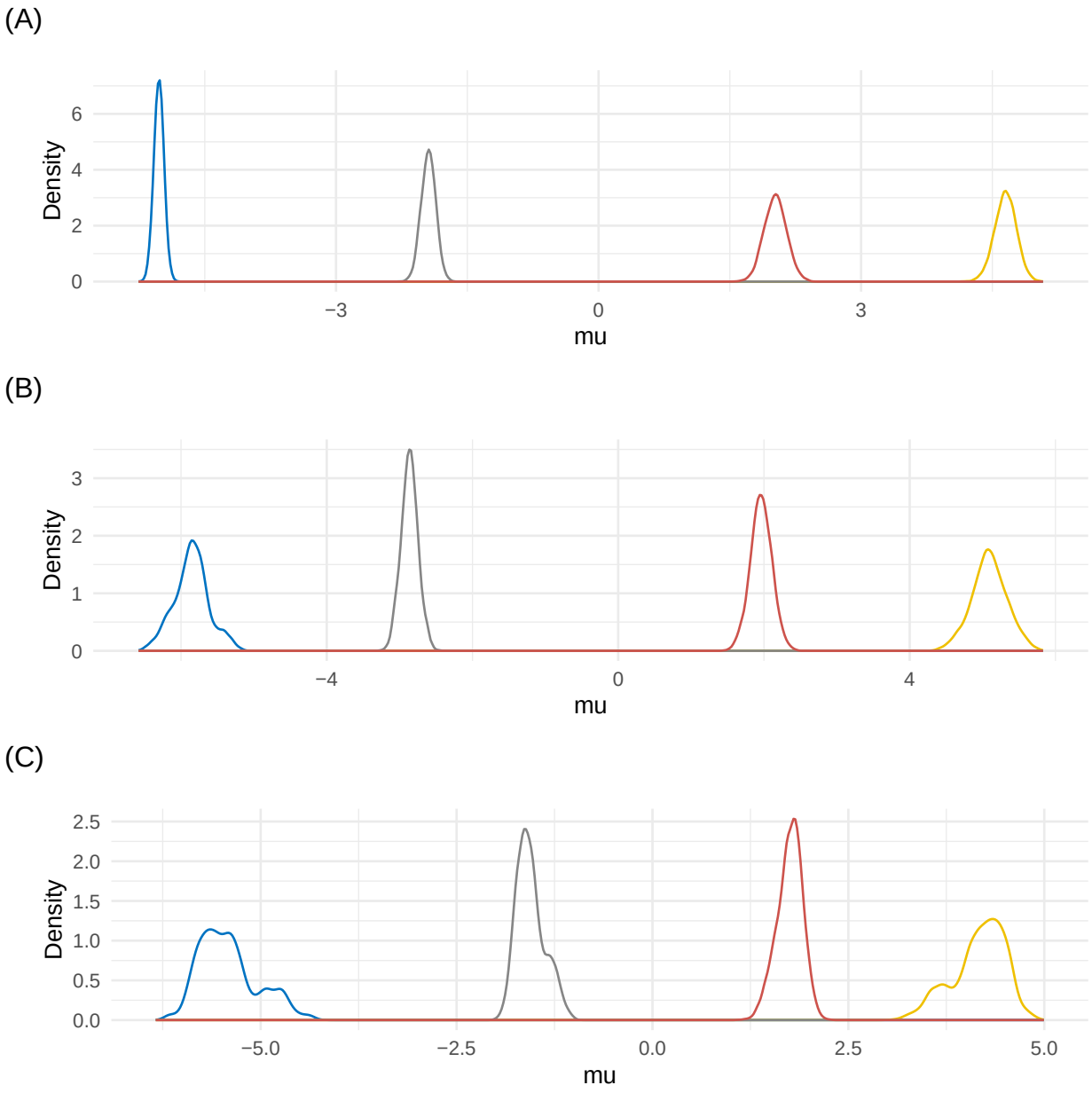
The sample of the MCMC is saved in `simBCFM`. To find the results of the simulated result, we generate the posterior distribution plot of the cluster assignment probabilities.

```
ggplot.probs.density(simBCFM)
```



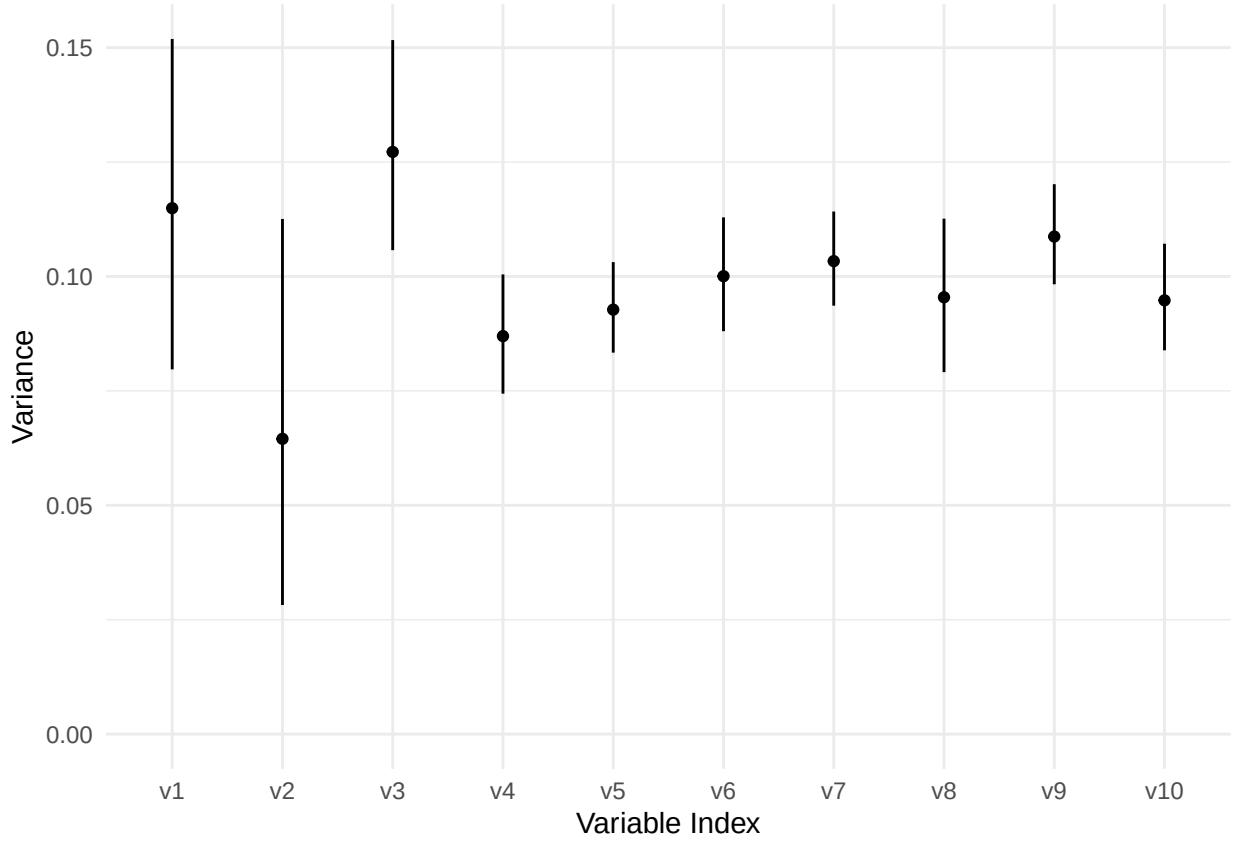
The plot below shows the posterior density of the cluster means.

```
ggplot.mu.density(simBCFM)
```



The plot below shows the credible intervals of the observational errors σ^2.

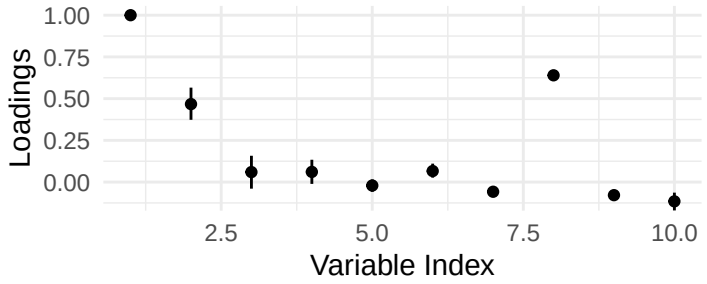
```
ggplot.sigma2.CI(simBCFM)
```



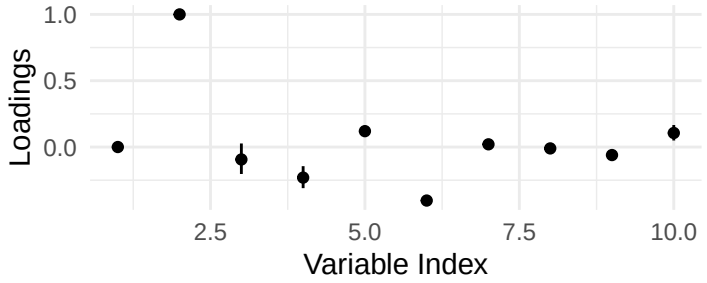
Figures below show the credible intervals of the factor loadings and the heatmap of the observations.

```
ggplot.B.CI(simBCFM)
```

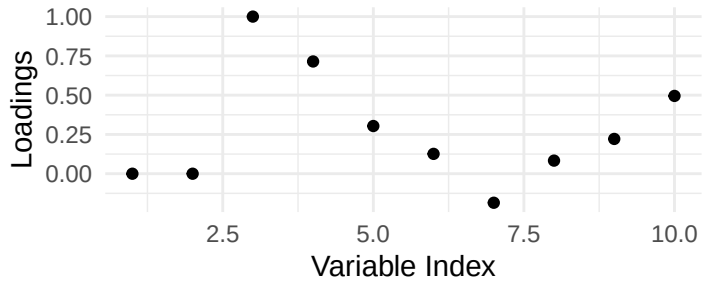
(A)



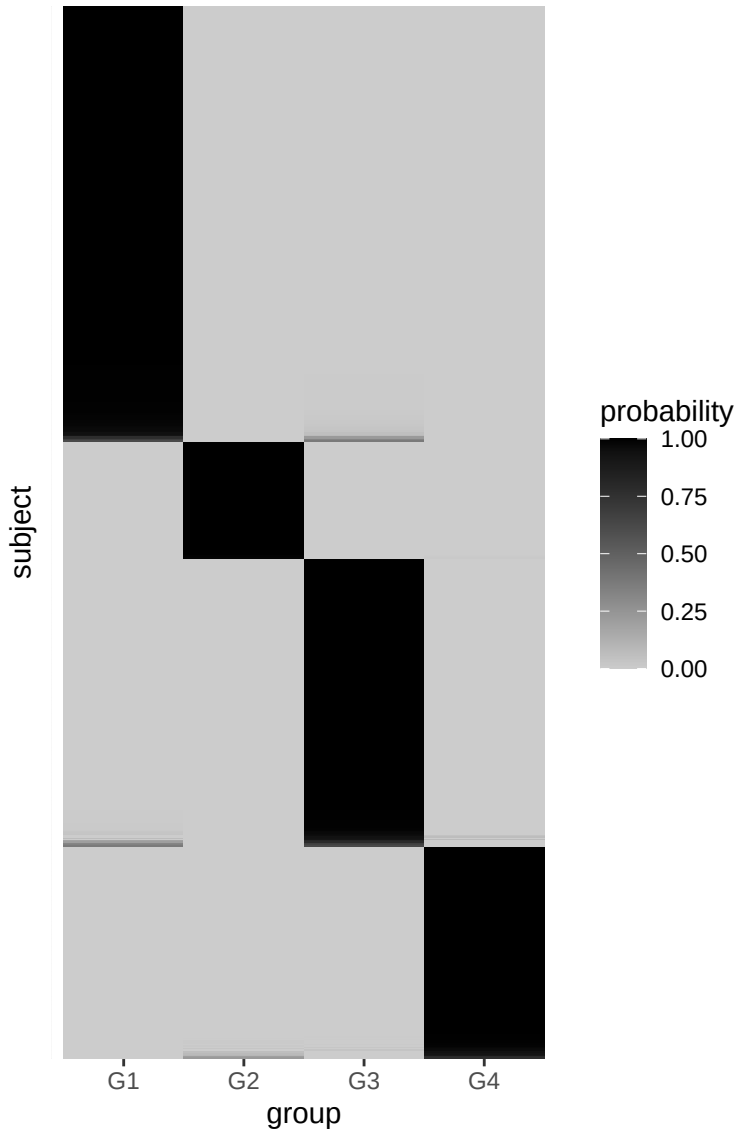
(B)



(C)




```
ggplot.Zit.heatmap(simBCFM)
```



In general, we run multiple model and evaluate to find the best setting. In this example, we assume the original number of clusters is 4 and number of factors is 3. We run models from 3 to 5 clusters and 2 to 4 factors and apply BIC with integrated likelihood. We first make a loop to generate MCMC sample with 3,000 iterations.

```
set.seed(1101)
n.iter <- 3000
simBCFM.list <- vector("list", length = 9)
currenti <- 0

for (nclusters in 3:5) {
    for (nfactors in 2:4) {
        cat("Started BCFM # of clusters =", nclusters, "# of factors =", nfactors,
            "\n")
        currenti <- currenti + 1
        simBCFM.current <- vector("list", length = 4)
        model.attributes <- initialize.model.attributes(S = 800, times = 1, R = 10,
```

```
            L = nfactors, G = nclusters)
        # cluster.hyperparms <- initialize.cluster.hyperparms(sim.data$data,
        # model.attributes) hyp.parm <- initialize.hyp.parm(model.attributes,
        # cluster.hyperparms, p.exponent = rep(2, nclusters))
        hyp.parm <- initialize.hyp.parm(data = data, model.attributes = model.attributes,
            p.exponent = rep(2, nclusters), vague.mu = TRUE)

        tic <- Sys.time()
        simBCFM <- BCFMcpp(sim.data$data, model.attributes, hyp.parm, n.iter, every = 10,
            verbose = FALSE)
        toc <- Sys.time()
        run.time <- toc - tic
        cat("Run time:", run.time, "\n")

        simBCFM.current[[1]] <- model.attributes
        simBCFM.current[[2]] <- hyp.parm
        simBCFM.current[[3]] <- simBCFM
        simBCFM.current[[4]] <- run.time
        names(simBCFM.current) <- c("model.attributes", "hyp.parm", "simBCFM", "run.time")
        simBCFM.list[[currenti]] <- simBCFM.current
    }
}
```

Now, we evaluate the model through BIC with integrated likelihood. The result of the 9 models is in the table below.

```
simBCFM.BIC <- matrix(NA, 3, 3)
for (i in 1:9) {
    simBCFM.BIC[i] <- BIC.like(sim.data$data, simBCFM.list[[i]]$simBCFM, simBCFM.list[[i]]$model.attribu
}
simBCFM.BIC <- as.data.frame(simBCFM.BIC)
rownames(simBCFM.BIC) <- paste("Factor =", 2:4)
colnames(simBCFM.BIC) <- paste("Cluster =", 3:5)
kable(simBCFM.BIC)
```

	Cluster = 3	Cluster = 4	Cluster = 5
Factor = 2	15596.17	15493.61	15520.38
Factor = 3	13751.88	13612.35	13642.64
Factor = 4	13866.58	13742.76	13823.57

In case of BIC with integrated likelihood, smaller value indicates better performance. From the information, we can find that the 4 clusters and 3 factors model has the best performance, which is the true setting of the simulated dataset.

Reference

- Prado, R., Ferreira, M. A. R., and West, M. (2021). Time Series: Modeling, Computation, and Inference. Boca Raton (2nd Ed.): Chapman & Hall/CRC.
- Aguilar, O. and West, M. (2000). "Bayesian dynamic factor models and portfolio allocation." Journal of Business and Economic Statistics, 18, 338–357.

- Lopes, H. and West, M. (2004). "Bayesian Model Assessment in Factor Analysis." Statistica Sinica, 14, 41–67.

Chapter 5

Conclusion and Future Research

5.1 Conclusion

In this dissertation, we proposed two novel classes of Bayesian factor models: Dynamic ICAR Spatiotemporal Factor Models (DIFM) and the Bayesian Clustering Factor Models (BCFM).

By incorporating spatiotemporal structures into a factor model, DIFM can account for the relationships from neighboring regions and autocorrelations from consecutive time observations. We can quantify the strength of spatial correlations and capture the behavior of the temporal components. DIFM imposes hierarchical structural constraint on the factor loadings matrix to ensure the model is identifiable and interpretable. We discovered the number of factors is significantly smaller than the dimension of the data in both the simulated dataset and case study. In addition, we can derive n-step-ahead forecasts to compute the predicted mean and intervals of each common factor. In the simulated dataset, we found the Laplace-Metropolis estimator of the predictive density selects the correct model. We applied DIFM to a dataset of the number of deaths due to drug overdose in United States. The criterion identified 9-factor model with the best performance. Also, our study found there are spatial correlations among adjacent states and common temporal trends of factors.

In BCFM, we detect latent clusters within the dataset by assuming Gaussian mixture model on the common factors. We first set the number of clusters and factors to launch BCFM. The initial number of factors can be determined through principal component analysis and exploratory factor analysis. To decide the initial number of clusters, we employ k-means clustering. This step is crucial for constructing hyperparameters of the informative priors of cluster assignment probabilities, means, and covariances. In addition to the hierarchical structural constraint, we impose an additional constraint on the first cluster. The covari-

ance matrix of the first cluster should be diagonal, and it represents the largest cluster in the initial step. This constraint prevents the model from label switching issues and contributes to convergence of parameters. In the simulated dataset, our BCFM found the correct setting in both evaluation criteria: Laplace-Metropolis estimator of the predictive density and BIC with integrated likelihood criterion. Across 300 simulated datasets with three different settings based on the distances among centroids, BIC with integrated likelihood criterion outperformed the Laplace-Metropolis estimator. We applied BCFM to the baseline of long-term recovery from opioid use disorder dataset [15]. We found 4 clusters and 3 factors were the most appropriate setting for both the Laplace-Metropolis estimator and the BIC with integrated likelihood criterion. This setting aligns with the results of previous research on this longitudinal dataset [14]. Finally, we applied BCFM to the breast cancer molecular subtype dataset. The BIC with integrated likelihood criterion chose 4 clusters and 15 factors, and the Laplace-Metropolis estimator selected 4 clusters and 19 factors model. Notably, the number of clusters matches the number of subtypes in the dataset.

We built two R packages: `DIFM` and `BCFM`. The packages allow researchers convenient access and application to the Bayesian factors models we propose in this dissertation. The `DIFM` package includes functions to run MCMC and visualize the posterior distribution of Dynamic Spatiotemporal Factor Models. Also, Package `BCFM` contains functions to simulate parameters from Gibbs sampling and output posterior density plots of Bayesian Clustering Factor Models.

5.2 Future Research

Bayesian factor models can be applied to a variety of research, including analyzing large-scale datasets. We developed methods to account for spatiotemporal correlations and to integrate clustering methods. The models we proposed are feasible and applicable to a wide range of fields of study. Our work can be extended to several avenues for future research.

In DIFM, we can consider numerous temporal trends within the data. In this dissertation, we used a second-order polynomial dynamic linear model (DLM) to address non-stationary behaviors of the observations. However, DLMs can also accommodate cases when the time

series is stationary, as well as cyclical and seasonal behaviors. This flexibility allows the model to capture various temporal patterns where different structures of the evolution matrix can be assumed.

Another utilization is considering different distributions for the observations. In Chapter 2, we discussed Gaussian distribution for both factor loadings and common factors. This is appropriate for continuous and large count data. However, when observations are discrete variables, researchers should use a different approach. For instance, in the case of small count data, it may be appropriate to assume that the observations follow a Poisson distribution for more accurate analysis.

Furthermore, we can explore different methods for setting the hyperparameters of the clusters. The posterior density of BCFM is especially sensitive to hyperaparameters, as it heavily depends on the size of the cluster covariance, means and assignment probabilities. We have applied principal component analysis and exploratory factor analysis followed by k-means to determine hyperparameters of the priors of cluster means and covariances. For cluster assignment hyperparameters, we used $(\alpha_1, \ldots, \alpha_K) = (2, \ldots, 2)$. This is a weakly informative prior that may facilitate faster convergence and avoid building empty clusters. However, depending on the characteristics of the datasets, these settings may not always be reasonable. When PCA and k-means struggle to find the latent clusters in the dataset, BCFM may also yield poor results. To resolve this problem, future research can focus on developing more effective methods for identifying initial clusters and factors.

Finally, we suggest additional studies in evaluating the models. One of the primary challenges in unsupervised clustering models is selecting the optimal number of clusters. In BCFM, we implemented two evaluation methods: the Laplace-Metropolis estimator of predictive density and the BIC with integrated likelihood. While the BIC with integrated likelihood was able to find the correct settings for moderately separated simulated datasets, the method tended to choose fewer clusters when the observations are not well-separated. Introducing a novel assessment method for determining the best model configuration would be a valuable extension of this research.

Appendices

Appendix A

Full Conditionals for Dynamic ICAR Spatiotemporal Factors Model

Full conditional distribution of idiosyncratic variance σ_j^2

The full conditional density of the jth idiosyncratic variance σ_j^2 can be obtained through

$$\begin{aligned} p(\sigma_j^2|-) &\propto f(\mathbf{y}_{.j}|\mathbf{X}, \mathbf{b}_j, \sigma_j^2) p(\sigma_j^2) \\ &\propto (\sigma_j^2)^{-n/2} exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n}(y_{ij} - \mathbf{x}_i\mathbf{b}_j)^2\right\}(\sigma_j^2)^{-n_\sigma/2} exp\left(-\frac{1}{2\sigma_j^2}n_\sigma s_\sigma^2\right) \\ &\propto (\sigma_j^2)^{-(n+n_\sigma)/2-1} exp\left[-\frac{1}{2\sigma_j^2}\left\{\sum_{i=1}^{n}(y_{ij} - \mathbf{x}_i\mathbf{b}_j)^2 + n_\sigma s_\sigma^2\right\}\right]. \end{aligned}$$

Therefore, the full conditional distribution of σ_j^2 is inverse gamma $IG((n+n_\sigma)/2, (\sum_{i=1}^{n}(y_{ij} - \mathbf{x}_i\mathbf{b}_j)^2 + n_\sigma s_\sigma^2)/2))$.

Full conditional distribution of spatial dependence parameter τ_j

Let $\mathbf{B}^*_{.j}$ be the vector of factor loadings of the jth factor without the first j fixed factor loadings, and $\mathbf{H}_j$ be obtained from the matrix $\mathbf{H}$ by removing its first j rows and j columns. Then the full conditional density of τ_j can be obtained as

$$\begin{aligned} p(\tau_j|-) &\propto p(\tau_j) p(\mathbf{B}^*_{.j}|\tau_j, \mathbf{H}_j, \mathbf{B}_{1:j,j} = (0, \ldots, 0, 1)^T) \\ &\propto (\tau_j)^{-(n_\tau + r - j)/2 - 1} exp\left\{-\frac{1}{2\tau_j}(n_\tau s_\tau^2 + (\widehat{\mathbf{B}}^*_{.j} - \mathbf{h}_j)^T \mathbf{H}^*_j (\widehat{\mathbf{B}}^*_{.j} - \mathbf{h}_j)\right\}, \end{aligned}$$

where $\mathbf{H}_j^* = \mathbf{H}_{(j+1):r,(j+1):r}$ and $\mathbf{h}_j = -\mathbf{H}_j^{*-1}\mathbf{H}_{(j+1):r,j}$ Therefore, the full conditional distribution of the spatial dependence parameter τ_j is inverse gamma $IG((n_\tau + r - j)/2, (n_\tau s_\tau^2 + \mathbf{B}_{\cdot j}^{*'}\mathbf{H}_j\mathbf{B}_{\cdot j}^*)/2)$.

Full conditional distribution of evolution covariance matrix W

The full conditional density of the evolution covariance matrix $\mathbf{W}$ is

$$\begin{aligned}
p(\mathbf{W}|-) &\propto p(\mathbf{W})p(\boldsymbol{\theta}|\mathbf{W},\mathbf{G}) \\
&\propto \prod_{t=2}^{n} |\mathbf{W}|^{-1/2} exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})'\mathbf{W}^{-1}(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})\right\} \\
&\quad |\mathbf{W}|^{-(n_W+2k+1)/2} exp\left\{-\frac{1}{2}tr(\mathbf{W}^{-1}\mathbf{S}_W)\right\} \\
&= |\mathbf{W}|^{-(n-1)/2} exp\left[-\frac{1}{2}\sum_{t=2}^{n} tr\{\mathbf{W}^{-1}(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})'\}\right] \\
&\quad |\mathbf{W}|^{-(n_W+2k+1)/2} exp\left\{-\frac{1}{2}tr(\mathbf{W}^{-1}\mathbf{S}_W)\right\} \\
&\propto \mathbf{W}^{-(n_W+2k+n)/2} exp\left[-\frac{1}{2}tr\left\{\mathbf{W}^{-1}(\mathbf{S}_W + \sum_{t=1}^{n}(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})'(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})\right\}\right].
\end{aligned}$$

Therefore, the full conditional is follows an inverse Wishart distribution $IW(\mathbf{S}_W + \sum_{t=2}^{n}(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})'(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1}), n_W + n - 1)$.

Appendix B

Full Conditionals for Bayesian Clustering Factors Model

Full conditional of factor loadings matrix B

Let $\mathbf{b}_r$ be the rth row of the factor loadings matrix $\mathbf{B}$. When $r > F$, the full conditional of $\mathbf{b}_r$ is

$$
\begin{aligned}
p(\mathbf{b}_r|-) &\propto f(\mathbf{Y}|\mathbf{X},\mathbf{B},\mathbf{V})p(\mathbf{b}_r) \\
&\propto exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{B}\mathbf{x}_i)^{'}\mathbf{V}^{-1}(\mathbf{y}_i-\mathbf{B}\mathbf{x}_i)\right)exp\left(-\frac{1}{2}\mathbf{b}_r^{'}\mathrm{T}^{-1}\mathbf{b}_r\right) \\
&\propto exp\left(-\frac{1}{2\sigma_r^2}(\mathbf{y}_{.,r}-\mathbf{X}\mathbf{b}_r)^{'}(\mathbf{y}_{.,r}-\mathbf{X}\mathbf{b}_r)\right)exp\left(-\frac{1}{2}\mathbf{b}_r^{'}\mathrm{T}^{-1}\mathbf{b}_r\right) \\
&\propto exp\left(-\frac{1}{2}\left(\mathbf{b}_r^{'}\mathrm{T}^{-1}\mathbf{b}_r+exp\left(-\frac{1}{2\sigma_r^2}(\mathbf{y}_{.,r}^{'}-\mathbf{b}_r^{'}\mathbf{X}^{'})(\mathbf{y}_{.,r}-\mathbf{X}^{'}\mathbf{b}_r)\right)\right)\right) \\
&\propto exp\left(-\frac{1}{2}\left(\mathbf{b}_r^{'}\left(\frac{1}{\sigma_r^2}\mathbf{X}^{'}\mathbf{X}+\mathrm{T}^{-1}\right)\mathbf{b}_r-\frac{2}{\sigma_r^2}\mathbf{b}_r^{'}\mathbf{X}^{'}\mathbf{y}_{.,r}\right)\right).
\end{aligned}
$$

Therefore, the full conditional of $\mathbf{b}_r$ $(r > F)$ is Gaussian distribution $N((\mathbf{X}^{'}\mathbf{X}/\sigma_j^2 + \mathrm{T}^{-1})^{-1}\mathbf{X}^{'}\mathbf{y}_{.,r}, (\mathbf{X}^{'}\mathbf{X}/\sigma_r^2+\mathrm{T}^{-1})^{-1}\mathbf{X}^{'}\mathbf{y}_{.,r})$

When $1 < r < F$, there are $F-r+1$ first fixed elements $\mathbf{b}_r$ due to the hierarchical structural constraint. The rth element is 1, and the first $r-1$ ($2 < r < F$ elements are set 0. Let $\mathbf{b}_r^*$ be the $\mathbf{b}_r$ with the last $r-1$ elements. Then the full conditional of $\mathbf{b}_r^*$ is

$$
\begin{aligned}
p(\mathbf{b}_r^*|-) &\propto f(\mathbf{Y}|\mathbf{X},\mathbf{B},\mathbf{V})p(\mathbf{b}_r^*) \\
&\propto exp\left(-\frac{1}{2\sigma_r^2}(\mathbf{y}_{.,r}^{'}-\mathbf{X}_{.,r}^{'}-\mathbf{b}_r^{*'}\mathbf{X}_{.,1:r-1}^{'})(\mathbf{y}_{.,r}-\mathbf{X}_{.,r}-\mathbf{X}_{.,1:r-1}\mathbf{b}_r^*)\right) exp\left(-\frac{1}{2}\mathbf{b}_r^{*'}\mathrm{T}_{1:r-1,1:r-1}^{*-1}\mathbf{b}_r^*\right) \\
&\propto exp\left(-\frac{1}{2}\left(\mathbf{b}_r^{*'}\left(\frac{1}{\sigma_r^2}\mathbf{X}_{.,1:r-1}^{'}\mathbf{X}_{1:r-1}+\mathrm{T}_{1:r-1,1:r-1}^{-1}\right)\mathbf{b}_r^*-2\frac{1}{\sigma_r^2}\mathbf{b}_r^{*'}\mathbf{X}_{.,1:r}^{'}(\mathbf{y}_{.,r}-\mathbf{X}_{.,r})\right)\right).
\end{aligned}
$$

Therefore, the full conditional of $\mathbf{b}_r^*$ is Gaussian distribution $N((\mathbf{X}_{.,r}^{'}\mathbf{X}_{.,r}/\sigma_r^2+\mathrm{T}_{1:r-1,1:r-1}^{-1})^{-1}\mathbf{X}_{.,1:r-1}(\mathbf{y}_{.,r}-\mathbf{X}_{.,r})/\sigma_r^2,(\mathbf{X}_{.,1:r-1}^{'}\mathbf{X}_{1:r-1}/\sigma_r^2+\mathrm{T}_{1:r-1,1:r-1}^{-1})^{-1})$

Full conditional of common factors matrix X

Let the group assignment of the ith observation to cluster k be $z_i = k$. Consier the case when the ith observation is assigned to the kth cluster. Then the full conditional of $\mathbf{x}_i$ can be derived through

$$\begin{aligned}
p(\mathbf{x}_i|-) &\propto p(\mathbf{x}_i)p(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k, z_i = k) \\
&\propto exp\left(-\frac{1}{2}\left((\mathbf{y}_i^{'} - \mathbf{x}_i^{'}\mathbf{B}^{'})\mathbf{V}^{-1}(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i) + (\mathbf{x}_i^{'} - \boldsymbol{\mu}_k^{'})\boldsymbol{\Omega}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right)\right) \\
&\propto exp\left(-\frac{1}{2}\left(\mathbf{x}_i^{'}\mathbf{B}^{'}\mathbf{V}^{-1}\mathbf{B}\mathbf{x}_i - 2\mathbf{x}_i^{'}\mathbf{B}^{'}\mathbf{V}^{-1}\mathbf{y}_i + \mathbf{x}_i^{'}\boldsymbol{\Omega}_k^{-1}\mathbf{x}_i - 2\mathbf{x}_i^{'}\boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k\right)\right) \\
&\propto exp\left(-\frac{1}{2}(\mathbf{x}_i^{'}(\mathbf{B}^{'}\mathbf{V}^{-1}\mathbf{B} + \boldsymbol{\Omega}_k^{-1})\mathbf{x}_i - 2\mathbf{x}_i^{'}(\mathbf{B}^{'}\mathbf{V}^{-1}\mathbf{y}_i + \boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k))\right).
\end{aligned}$$

Therefore, the full conditional of $\mathbf{x}_i$ is Gaussian distribution $N((\boldsymbol{\Omega}_k^{-1}+\mathbf{B}^{'}\mathbf{V}^{-1}\mathbf{B})^{-1}(\mathbf{B}^{'}\mathbf{V}^{-1}\mathbf{y}_i+ \boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k), (\boldsymbol{\Omega}_k^{-1} + \mathbf{B}^{'}\mathbf{V}^{-1}\mathbf{B})^{-1})$.

Full conditional of the kth cluster mean vector $\boldsymbol{\mu}_k$

Let n_k be the cardinality of C_k. Then, the full conditional density of $\boldsymbol{\mu}_k$ is

$$\begin{aligned}
p(\boldsymbol{\mu}_k|-) &\propto p(\boldsymbol{\mu}_k) \prod_{i \in C_k} p(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k, z_i = k) \\
&\propto exp\left(-\frac{1}{2}\sum_{i \in C_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)^{'}\boldsymbol{\Omega}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right) exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\right)^{'}\boldsymbol{\Omega}_k^{*-1}\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\right)\right) \\
&\propto exp\left(-\frac{1}{2}\sum_{i \in C_k}(\boldsymbol{\mu}_k^{'}\boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k\boldsymbol{\Omega}_k^{-1}\mathbf{x}_i)\right) exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}_k^{'}\boldsymbol{\Omega}_k^{*-1}\boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k\boldsymbol{\Omega}_k^{*-1}\boldsymbol{\mu}_k^*\right)\right) \\
&\propto exp\left(-\frac{1}{2}\boldsymbol{\mu}_k^{'}\left(n_k\boldsymbol{\Omega}_k^{-1} + \boldsymbol{\Omega}_k^{*-1}\right)\boldsymbol{\mu}_k - \boldsymbol{\mu}_k\left(\boldsymbol{\Omega}_k^{-1}\sum_{i \in C_k}\mathbf{x}_i + \boldsymbol{\Omega}_k^{*-1}\boldsymbol{\mu}_k^*\right)\right).
\end{aligned}$$

Therefore, the full conditional of $\boldsymbol{\mu}_k$ is Gaussian distribution
$N((\boldsymbol{\Omega}_k^{*-1} + n_k\boldsymbol{\Omega}_k^{-1})^{-1}\left(\boldsymbol{\Omega}_k^{*-1}\boldsymbol{\mu}_k^* + n_k\boldsymbol{\Omega}_k^{-1}\bar{\mathbf{X}}_k^*\right), \left(\boldsymbol{\Omega}_k^{*-1} + n_k\boldsymbol{\Omega}_k^{-1}\right)^{-1})$, where $\bar{\mathbf{X}}_k^* = \sum_{i \in C_k}\mathbf{x}_i/n_k$.

Full conditional of the kth cluster covariance $\mathbf{\Omega}_k$

The covariance matrix of the first cluster $\mathbf{\Omega}_1$ is diagonal, thus we apply each of diagonal elements inverse gamma prior. Let the lth diagonal value ($l = 1, \ldots, F$) of $\mathbf{\Omega}_1$ be ω_{1l}. Then the full conditional is

$$\begin{aligned} p(\omega_{1l}|-) &\propto p(\omega_{1l}) \prod_{i \in C_1} p(\mathbf{x}_i|\boldsymbol{\mu}_1, \omega_{1l}, z_i = 1) \\ &\propto \omega_{1l}^{-n_1/2-n_\omega/2-1} exp\left(-\frac{1}{2\omega_{1l}}\left(\sum_{i \in C_1}(x_{il} - \mu_{1l})^2 + n_\omega s_\omega^2\right)\right), \end{aligned}$$

Therefore, the full conditional distribution of ω_{1l} is inverse gamma distribution $IG((n_1 + n_\omega)/2, (\sum_{i \in C_1}(x_{il} - \mu_{1l})^2 + n_\omega s_\omega^2)/2)$

The cluster covariance $\mathbf{\Omega}_k$ when $k > 1$ do not have constraints. Therefore, we apply inverse Wishart prior. The full conditional density of $\mathbf{\Omega}_k$ ($k > 1$) is

$$\begin{aligned} p(\mathbf{\Omega}_k|\_) &\propto p(\mathbf{\Omega}_k) \prod_{i \in C_k} p(\mathbf{x}_i|\boldsymbol{\mu}_k, \mathbf{\Omega}_k, z_i = k) \\ &\propto |\mathbf{\Omega}_k|^{-(n_k+\nu+F+1)/2} exp\left(-\frac{1}{2}\left(tr\left(\mathbf{\Omega}_k^{-1}\sum_{i \in C_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)'\right) + tr(\mathbf{\Omega}_k^{-1}\mathbf{\Psi}_k)\right)\right) \\ &\propto |\mathbf{\Omega}_k|^{-(n_k+\nu+F+1)/2} exp\left(-\frac{1}{2}\left(tr\left(\mathbf{\Omega}_k^{-1}\left(\sum_{i \in C_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)' + \mathbf{\Psi}_k\right)\right)\right)\right) \end{aligned}$$

The full conditional distribution of $\mathbf{\Omega}_k$ is inverse Wishart distribution $IW(n_k + \nu, \sum_{i \in C_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)' + \mathbf{\Psi}_k)$.

Full conditional of factor loadings variance τ_l

Let $\mathbf{b}_l^*$ be the lth column of $\mathbf{B}$ without the first l elements. Then, the full conditional distribution of τ_l can be obtained by

$$\begin{aligned} p(\tau_l|-) &\propto p(\tau_l)p(\mathbf{b}_l^*|\tau_l) \\ &\propto \frac{n_\tau s_\tau^2/2}{\Gamma(n_\tau/2)}\tau_l^{-(n_\tau/2+1)}exp\left(-\frac{n_\tau s_\tau^2/2}{\tau_l}\right)\tau_l^{-(R-l)/2}exp\left(-\frac{1}{2\tau_l}\mathbf{b}_l^{*'}\mathbf{b}_i^*\right) \\ &\propto \tau_l^{-(R-l+n_\tau)/2-1}exp\left(-\frac{1}{2\tau_l}(\mathbf{b}_l^{*'}\mathbf{b}_l^* + n_\tau s_\tau^2)\right). \end{aligned}$$

Therefore, the full conditional of τ_l is inverse gamma distribution $IG((R-l+n_\tau)/2, (\mathbf{b}_l^{*'}\mathbf{b}_l^* + n_\tau s_\tau^2)/2)$.

Full conditional of idiosyncratic variance σ_r^2

The full conditional of the rth idiosyncratic variance σ_r^2 can be obtained by

$$\begin{aligned} p(\sigma_r^2|\_) &\propto p(\sigma_r^2)f(\mathbf{y}_{\cdot r}|\mathbf{X}, \mathbf{b}_r, \sigma_r^2) \\ &\propto (\sigma_r^2)^{-n/2}exp\left(-\frac{1}{2\sigma_r^2}\sum_{i=1}^{n}(y_{ir} - \mathbf{x}_i\mathbf{b}_r)^2\right) \\ &\propto (\sigma_r^2)^{-(n+n_\sigma)/2-1}exp\left(-\frac{1}{2\sigma_r^2}\left(\sum_{i=1}^{n}(y_{ir} - \mathbf{x}_i\mathbf{b}_r)^2 + n_\sigma s_\sigma^2\right)\right). \end{aligned}$$

Therefore, the full conditional of σ_r^2 is $IG((n+n_\sigma)/2, (\sum_{i=1}^{n}(y_{ir} - \mathbf{x}_i\mathbf{b}_r)^2 + n_\sigma s_\sigma^2)/2))$

Full conditional of probabilities $p_1, \ldots, p_K$

Let the vector of cluster assignment probabilities $\mathbf{p} = (p_1, \ldots, p_K)$.The full conditional of the cluster probabilities $p_1, \ldots, p_K$ can be obtained through

$$\begin{aligned} p(\mathbf{p}|-) &\propto p(\mathbf{p})p(\mathbf{Z}|\mathbf{p}) \\ &\propto \prod_{k=1}^{K} p_k^{n_k} \prod_{k=1}^{K} p_k^{\alpha_k} \\ &\propto \prod_{k=1}^{K} p_k^{n_k+\alpha_k} \end{aligned}$$

Therefore, the full conditional of $\mathbf{p}$ is Dirichlet distribution $Dirichlet(n_1+\alpha_1, \ldots, n_K+\alpha_K)$.

Bibliography

[1] Aguilar, O., Huerta, G., Prado, R., and West, M. (1999). "Bayesian inference on latent structure in time series (with discussion)." In *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 3–26. Oxford University Press.

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

[3] Baek, J. and McLachlan, G. J. (2011). "Mixtures of common t-factor analyzers for clustering high-dimensional microarray data." *Bioinformatics*, 27, 9, 1269–1276.

[4] Ball, G. H., Hall, D. J., et al. (1965). *ISODATA, a novel method of data analysis and pattern classification*, vol. 699616. Stanford research institute Menlo Park, CA.

[5] Besag, J. and Kooperberg, C. (1995). "On conditional and intrinsic autoregression." *Biometrika*, 82, 4, 733–746.

[6] Besag, J., York, J., and Mollie, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43, 1.

[7] Bhattacharya, A. and Dunson, D. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98, 291–306.

[8] Bivand, R. and Wong, D. W. S. (2018). "Comparing implementations of global and local indicators of spatial association." *TEST*, 27, 3, 716–748.

[9] Capdeville, V., Gonçalves, K. C., and Pereira, J. B. (2021). "Bayesian factor models for multivariate categorical data obtained from questionnaires." *Journal of Applied Statistics*, 48, 16, 3150–3173.

[10] Carter, C. K. and Kohn, R. (1994). "On Gibbs sampling for state space models." *Biometrika*, 81, 541–553.

[11] Carvalho, C. M., Lopes, H. F., and Aguilar, O. (2011). "Dynamic stock selection strategies: A structured factor model framework." In *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 69–90. Oxford University Press.

[12] Chen, Q., Han, R., Ye, F., and Li, W. (2011). "Spatio-temporal ecological models." *Ecological Informatics*, 6, 1, 37–43. Special Issue: 5th Anniversary.

[13] Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., and Mathieu, C. (2017). "Hierarchical Clustering: Objective Functions and Algorithms." *CoRR*, abs/1704.02147.

[14] Craft, W. H., Shin, H., Tegge, A. N., Keith, D. R., Athamneh, L. N., Stein, J. S., Ferreira, M. A. R., Chilcoat, H. D., Le Moigne, A., DeVeaugh-Geiss, A., and Bickel, W. K. (2022). "Long-term recovery from opioid use disorder: recovery subgroups, transition states and their association with substance use, treatment and quality of life." *Addiction*, 118, 890–900.

[15] Craft, W. H., Tegge, A. N., Keith, D. R., Shin, H., Williams, J., Athamneh, L. N., Stein, J. S., Chilcoat, H. D., Le Moigne, A., DeVeaugh-Geiss, A., and Bickel, W. K. (2022). "Recovery from opioid use disorder: A 4-year post-clinical trial outcomes study." *Drug and Alcohol Dependence*, 234, 109389.

[16] Dama, F. and Sinoquet, C. (2021). "Analysis and modeling to forecast in time series: a systematic review." *CoRR*, abs/2104.00164.

[17] Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Ucar, I., Bates, D., and Chambers, J. (2023). *Rcpp: Seamless R and C++ Integration*. R package version 1.0.11.

[18] Eddelbuettel, D., Francois, R., Bates, D., Ni, B., and Sanderson, C. (2023). *RcppArmadillo: 'Rcpp' Integration for the 'Armadillo' Templated Linear Algebra Library*. R package version 0.12.6.4.0.

[19] Frühwirth-Schnatter, S. (1994). "Data augmentation and dynamic linear models." *Journal of Time Series Analysis*, 15, 2, 183–202.

[20] Gelfand, A. E. and Banerjee, S. (2017). "Bayesian Modeling and Analysis of Geostatistical Data." *Annual Review of Statistics and Its Application*, 4, 1, 245–266.

[21] Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85, 410, 398–409.

[22] — (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85, 410, 398–409.

[23] Geweke, J. and Zhou, G. (1996). "Measuring the price of the Arbitrage Pricing Theory." *The Review of Financial Studies*, 9, 2, 557–587.

[24] Guttman, L. (1954). "Some necessary conditions for common-factor analysis." *Psychometrika*, 19, 2, 149–161.

[25] Hogan, J. W. and Tchernis, R. (2004). "Bayesian Factor Analysis for Spatially Correlated Data, with Application to Summarizing Area-Level Material Deprivation from Census Data." *Journal of the American Statistical Association*, 99, 466, 314–324.

[26] Kaiser, H. F. (1960). "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement*, 20, 1, 141–151.

[27] Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the american statistical association*, 90, 431, 928–934.

[28] Keefe, M. J., Ferreira, M. A. R., and Franck, C. T. (2018). "On the formal specification of sum-zero constrained intrinsic conditional autoregressive models." *Spatial Statistics*, 24, 54–65.

[29] — (2019). "Objective Bayesian analysis for Gaussian hierarchical models with intrinsic conditional autoregressive priors." *Bayesian Analysis*, 14, 181 – 209.

[30] Knorr-Held, L. (2000). "Bayesian modelling of inseparable space-time variation in disease risk." *Statistics in Medicine*, 19, 17-18, 2555–2567.

[31] Kumar, A. and Gupta, S. C. (2015). "A new Initial Centroid finding Method based on Dissimilarity Tree for K-means Algorithm."

[32] Lee, D., Rushworth, A., and Napier, G. (2018). "Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package." *Journal of Statistical Software*, 84, 9, 1–39.

[33] Leorato, S. and Mezzetti, M. (2021). "A Bayesian Factor Model for Spatial Panel Data with a Separable Covariance Approach." *Bayesian Analysis*, 16, 2, 489 – 519.

[34] Lewis, S. M. and Raftery, A. E. (1997). "Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator." *Journal of the American Statistical Association*, 92, 438, 648–655.

[35] Lopes, H., Gamerman, D., and Salazar, E. (2011). "Generalized spatial dynamic factor models." *Computational Statistics & Data Analysis*, 55, 1319–1330.

[36] Lopes, H. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14, 41–67.

[37] Lopes, H. F., Salazar, E., and Gamerman, D. (2008). "Spatial dynamic factor analysis." *Bayesian Analysis*, 3, 4, 759–792.

[38] Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14, 41–67.

[39] Lòpez-Abente, G., Aragonés, N., García-Pérez, J., and Fernández-Navarro, P. (2014). "Disease mapping and spatio-temporal analysis: importance of expected-case computation criteria." *Geospatial Health*, 9, 1, 27–35.

[40] Murphy, K., Viroli, C., and Gormley, I. C. (2020). "Infinite Mixtures of Infinite Factor Analysers." *Bayesian Analysis*, 15, 3, 937 – 963.

[41] Murtagh, F. and Contreras, P. (2012). "Algorithms for hierarchical clustering: an overview." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2.

[42] Osorio, F. and Ogueda, A. (2024). *Fast computation of some matrices useful in statistics*. R package version 0.5-772.

[43] Paciorek, C. J. (2013). "Spatial models for point and areal data using Markov random fields on a fine grid." *Electronic Journal of Statistics*, 7, none, 946 – 972.

[44] Papastamoulis, P. (2020). "Clustering multivariate data using factor analytic Bayesian mixtures with an unknown number of components." *Statistics and Computing*, 30, 485–506.

[45] Papastamoulis, P. and Ntzoufras, I. (2022). "On the identifiability of Bayesian factor analytic models." *Statistics and Computing*, 32, 2, 1–29.

[46] Pebesma, E. J. and Bivand, R. S. (2005). "Classes and methods for spatial data in R." *R News*, 5, 2, 9–13.

[47] Prado, R., Ferreira, M. A. R., and West, M. (2021). *Time Series: Modeling, Computation, and Inference 2nd Ed.*. Boca Raton: Chapman & Hall/CRC.

[48] R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[49] Roth, R. E., Ross, K. S., Finch, B. G., Luo, W., and MacEachren, A. M. (2013). "Spatiotemporal crime analysis in U.S. law enforcement agencies: Current practices and unmet needs." *Government Information Quarterly*, 30, 3, 226–240.

[50] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Boca Raton, FL: Chapman and Hall.

[51] Shin, H. and Ferreira, M. A. (2023). "Dynamic ICAR Spatiotemporal Factor Models." *Spatial Statistics*, 56, 100763.

[52] Statisticat and LLC. (2021). *LaplacesDemon: Complete Environment for Bayesian Inference*. R package version 16.1.6.

[53] Stephens, M. (2002). "Dealing With Label Switching in Mixture Models." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62, 4, 795–809.

[54] Sun, J., Warren, L., and Zhao, H. (2016). "A Bayesian Semiparametric Factor Analysis Model for Subtype Identification." *Statistical Applications in Genetics and Molecular Biology*, 16, 145–158.

[55] Vermunt, J. and Magidson, J. (2002). *Latent Class Cluster Analyses.*

[56] Vitor Capdeville, K. C. M. G. and Pereira, J. B. M. (2021). "Bayesian factor models for multivariate categorical data obtained from questionnaires." *Journal of Applied Statistics*, 48, 16, 3150–3173. PMID: 35707256.

[57] West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." In *Bayesian Statistics 7*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 723–732. Oxford University Press.

[58] West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (2nd Ed.)*. Berlin, Heidelberg: Springer-Verlag.

[59] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.