UNIVERSITA' COMMERCIALE “LUIGI BOCCONI”

PhD SCHOOL

PhD program in Statistics

Cycle: XXXIII

Disciplinary Field: SECS-S/01

Scalable Bayesian dynamic regression in neuroimaging

Advisor: Sonia PETRONE

Co-Advisor: Michele GUINDANI

PhD Thesis by

Wei ZHANG

ID number: 3004191

Year 2022

Abstract

The thesis is motivated by the study of brain effective connectivity using neuroimaging data, in particular, functional magnetic resonance imaging (fMRI) data and electroencephalography (EEG) data. We focus on a largely applied methodology to study effective connectivity, the vector autoregressive (VAR) model, as it is closely related to the notion of Granger causality. Statistical challenges in inference with VAR models include the high dimension of the parameter space and the choice of the number of lags. We address these challenges and propose a novel framework based on tensor decomposition to achieve dimension reduction. We adopt a Bayesian approach, which allows to incorporate information from experts and to give a formal quantification of uncertainty. We first develop a (static) Bayesian tensor VAR model with a careful choice of the prior distributions. However, the main objective of the thesis is to develop *dynamic* tensor VAR models, in order to take into account dynamic changing patterns of the brain connectivity and nonlinearities. The thesis thus contributes to the established and still growing literature on dynamics in brain activities.

We propose a Bayesian time-varying tensor VAR model that employs a tensor decomposition for the VAR coefficient matrices at different lags. Dynamically varying connectivity patterns are captured by assuming a latent binary state process that selects the active components of the tensor decomposition at each time via a novel Ising prior specification in the time domain, and we use carefully designed sparsity-inducing priors that allow to ascertain model complexity through the posterior distribution. The model is studied on

synthetic data and in a real fMRI study involving a book reading experiment.

We further explore a more direct specification of a time-varying tensor VAR model through dynamic shrinkage priors. While the above Ising prior specification essentially assumes transition in terms of discrete latent states, an alternative approach is to envisage smoother temporal transitions by modeling the time-varying coefficients as an autoregressive process. We pursue this approach with the additional objectives of dimension reduction and temporal dependent sparsity. Our contribution is to employ dynamic shrinkage priors, recently proposed for dynamic variable selection in a regression setting, for time-varying tensor VAR models. More specifically, we employ the dynamic spike and slab prior and the dynamic shrinkage process to define hierarchical Bayesian time-varying tensor VAR models for multiple homogeneous trials.

As an ongoing project, we aim to contribute to Bayesian statistical methodology for dynamic regression with multivariate time series by proposing a new process prior that has the generalized double Pareto (GDP) prior as the marginal distribution.

Acknowledgement

I will remain forever indebted to my advisors, Professor Sonia Petrone and Professor Michele Guindani, for their dedication, diligence and patience in helping me achieve this important milestone of my life. Our coauthor Professor Ivor Cribben provided valuable insights into the data.

To my family members, with whom the journey began and continues.

Contents

Abstract **i**

Acknowledgement **iii**

1 Introduction **1**

1.1 Introduction and thesis outline 1

1.2 Brain effective connectivity 5

1.3 Statistical challenges 11

1.4 Structure of the thesis 12

2 Bayesian Vector Autoregressive Models **15**

2.1 VAR model 16

2.2 Bayesian VAR model 20

2.3 Bayesian tensor VAR (BTVAR) model 21

2.4 Simulation studies 27

3 Bayesian Time-Varying Tensor Vector Autoregressive Models for Dynamic Effective Connectivity **29**

3.1 Introduction 31

3.2 Time-Varying Tensor VAR (TVT-VAR) model 33

3.3 Prior specifications 38

3.3.1 Ising prior on temporal transitions 38

3.3.2 Sparsity-inducing priors 43

3.3.3 Rank of the PARAFAC decomposition 44

3.3.4 Posterior Computation 45

3.4 Simulation Studies 47

3.5 Real Data Application 53

3.6 Discussion and future work 63

4 Dynamic Shrinkage Priors in Time-Varying Vector Autoregressive Models 77

4.1 Introduction 78

4.2 Time-varying Vector Autoregressive Models 81

4.3 Dynamic Shrinkage Priors 85

4.3.1 DSS process priors in BHTVT-VAR models 86

4.3.2 DSP priors in BHTVT-VAR models 89

4.4 Posterior Inferences 90

4.5 Simulation Studies 93

4.6 Real Data Application 96

5 Dynamic Generalized Double Pareto Process Priors 103

5.1 A Gumbel AR(1) process 106

5.2 Dependence between shrinkage profiles 109

5.3 Posterior Inferences 110

References 113

Chapter 1

Introduction

1.1 Introduction and thesis outline

In this thesis, we are motivated by the study of brain effective connectivity using neuroimaging data and in particular functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) data. Brain effective connectivity measures directed influences one brain region has on others, which is of special importance as it gives us insights on the role of each brain region as well as how brain as a whole functions when a person is in a resting state or when he executes some tasks (Stephan and Friston, 2010; Friston, 2011; Razi and Friston, 2016). Many approaches for brain effective connectivity exist such as dynamic causal modeling (DCM) first introduced in Friston et al. (2003) and structural equation modeling (SEM) (Hoyle, 1995; Ullman and Bentler, 2003). DCM is general in the sense that it encompasses most effective connectivity models but it is also computationally expensive (Daunizeau et al., 2011). Simplifying DCM is desired as it helps to achieve a good balance between model generality and practical usability. Vector Autoregressive (VAR) models can be regarded as one way of simplification and are widely applied to study brain effective connectivity due to their connection with the notion of Granger causality. However, there are relevant statistical challenges in inferences with

VAR models. For instance, the parameter space is quadratic in the time series dimension, invalidating conventional estimation methods like the ordinary least squares (OLS). The huge parameter space also leads to unstable parameter estimation. Another challenge is in model specification as to determine the order (number of lags) to be included in the model. Including too few lags leads to the risk of neglecting important past observations that have significant explanatory power over current observations; on the other hand, including unnecessary lags to the model would cause overfitting and excessive computation burden.

We address these challenges and propose a novel framework incorporating tensor and tensor decomposition to achieve dimension reduction. In spite of this, the reduced parameter dimension is still considerably large, and further methods are needed to induce shrinkage and sparsity. We adopt the Bayesian approach, which allows to incorporate information from experts and give a formal quantification of uncertainty (or partial information), as learning is expressed through conditional distributions. This is reflected in our thoughtful choice of a global-local *shrinkage prior* distribution that has an increasing shrinkage component which enables the data driven selection of the optimal lags in the VAR. The proposed Bayesian tensor VAR model is presented in the first chapter of the thesis. Here, the model's parameters are constant over time, and we refer to the model as "static".

In fact, the main objective of the thesis is to develop methodological tools for *dynamic* brain effective connectivity, that can capture dynamic changing patterns of the brain activity and of (directed) causal relationships.

There is well-established and still lively growing literature that accounts for dynamics in brain activities (Hutchison et al., 2013; Taghia et al., 2017; Park et al., 2018a; Warnick et al., 2018; Park et al., 2018b; Zarghami and Friston, 2020a). Such consideration poses a further challenge when we add a temporal dimension to the analysis. In light of this, a time-varying regression structure is required. The starting point in our study is the time-varying VAR model, which offers more flexibility than the static VAR model in accounting for possible dynamic brain effective connectivity. In the proposed Bayesian dynamic tensor VAR model (Chapter 3), the time-varying coefficients of the VAR are stacked into a tensor so that tensor decomposition can be employed to exploit the underlying lower dimensional structure. More specifically, we envisage that each component of the tensor decomposition is stochastically 'activated' though a latent binary time series, which has a Markovian temporal dependence defined through a novel, parsimonious Ising prior in the time domain, which encourages no activation at each time point as well as joint activation of two consecutive time points. The remaining components of the prior specification leverages on what we have developed for the static VAR model (Chapter 2). After we test our method's performance through simulation studies, we apply the approach to an fMRI dataset where signals for experiment subjects are measured while they read novels. The real data application delivers some interesting findings that reveal changing information flows between brain regions when it comes to plot twists.

In the above model, the temporal evolution essentially assumes transitions in terms of discrete latent states; consequently, the model is designed to adapt to abrupt changes, since it envisages that, at any time point, a state can either be active or inactive. If the objective is to have smoother temporal transitions, a natural solution could be model the time-varying coefficients more directly, as an autoregressive process. In Chapter 4 of the thesis, we follow this path while keeping the overall framework that applies tensor decomposition to the VAR coefficient matrix, and with the additional objectives of dimension reduction and temporal dependent sparsity. Indeed, we notice that there is a very recent and growing literature on dynamic shrinkage prior laws (Kalli and Griffin, 2014; Kowal et al., 2019; Rockova and McAlinn, 2021; Irie, 2019). These priors are proposed for dynamic variable selection in a regression setting. A novel contribution in Chapter 4 of the thesis is to employ dynamic shrinkage priors in the framework of time-varying tensor VAR models. More specifically, we employ the dynamic spike and slab prior and the dynamic shrinkage process to define hierarchical Bayesian time-varying tensor VAR models for multiple homogeneous trials.

Inspired by the time-varying VAR model, we contemplate on a general goal to develop Bayesian *dynamic* regression tools for multivariate time series analysis. The Bayesian *dynamic* regression that we refer to has not only time-varying parameters but also time-varying sparsity which assumes that the sparsity induced on time varying parameters is temporal dependent; that is, if at some point the shrinkage imposed on a parameter is strong then similar strength of shrinkage should be imposed on parameters at the following time points; this is reasonable prior knowledge. Many time-varying priors have been proposed in literature as discussed above. Usually these priors yield stationary marginal distributions for the processes (Kalli and Griffin, 2014; Kowal et al., 2019; Rockova and McAlinn, 2021; Irie, 2019), with the exception that the marginal distribution of the dynamic shrinkage processes proposed by Kowal et al. (2019) does not have an analytical form. We consider an analytically tractable marginal distribution of a time-varying prior to be highly desirable property as this gives us a clear idea of the shrinkage behavior our proposed prior induces. An ongoing project is to contribute to the growing body of literature on dynamic shrinkage priors by proposing a new process prior that has the generalized double Pareto (GDP) prior as the marginal distribution. The reasons why we focus on the GDP prior include: 1) the GDP prior has been well studies, for instance, it has a spike around 0 as the Laplace distribution while it also behaves similarly to the student's t distribution in the tails; 2) the GDP prior was proposed in the tensor regression context, and extending it to the dynamic case naturally gives rise to a new application to the dynamic tensor regression setting.

1.2 Brain effective connectivity

Understanding how human brains work has always been a research topic that attracts wide attention. Nowadays two distinct perspectives exist in the neuroscience community. One branch, termed *functional segregation*, builds upon the hypothesis that different functions are processed in different regions of the brain locally and separately in the absence of communication. Research developed under this framework focuses on activation in human brain images and has yielded important results that facilitate our understanding of brain functions. For example, hippocampus has been identified to be responsible for memory formation; damages to hippocampus are linked with Alzheimer's Disease (Hyman et al., 1984; West et al., 1994; Jin et al., 2004). In addition to understanding disease mechanism and disease diagnosis, another research topic that falls inside this realm is to estimate task-related activation patterns. In such a experiment, the brain signals of a single subject or multiple subjects are measured after certain experiment condition and they are used to detect active brain regions subsequent to that condition. Classical methods in this area include for instance, the generalized linear model (GLM), which states that

$$\mathbf{y}_n = X_n \boldsymbol{\beta}_n + \boldsymbol{\epsilon}_n.$$

The response $\mathbf{y}_n$ is measure brain signals time series of length T, the design matrix $X_n \in \mathbb{R}^{T \times p}$ contains experiment tasks, input stimuli and covariates and $\boldsymbol{\beta}_n \in \mathbb{R}^p$ are the coefficients associated with each of the independent variables. $n = 1, \ldots, N$ is the index of regions of interest (ROIs) and the GLM is estimated separatedly for each region (Gössl et al., 2001; Trujillo-Barreto et al., 2004; Flandin and Penny, 2007; Trujillo-Barreto et al., 2008). Despite of its tremendous success in explaining some human brain functions, this perspective neglects another vital aspect, the connection that indicates simultaneous fluctuation or casual information transmission, revealing the true underlying mechanisms that drive brain function. Therefore, a growing number of publications in the field turn to *functional integration* to address more complex phenomena in interdisciplinary domains such as neuro-economics. It is documented in Konovalov and Krajbich (2019) that the brain has two learning systems, namely model-free and model-based systems and they are associated with white matter pathways between the striatum and premotor cortex and between the striatum and ventromedial prefrontal cortex (vmPFC) respectively, providing further evidence that segregated regions of brain usually only specialize in certain functions. However, the perception of sophisticated economic and psychological constructs such as value involves both learning systems, which requires the interactions between the two regions of the brain.

Brain integration can be further divided into *functional connectivity* and *effective connectivity*, modeling human brain images from two angles. The former, functional connectivity, assumes undirected graphs whereas the latter, effective connectivity, imposes directed edges between nodes of the brain, referring to the information flow from one region to another (Friston, 1994, 2011). The distinction in the assumption touches two fundamentally different guidelines. Functional connectivity describes the dependence structure of observed brain images while effective connectivity finds mechanism to explain the observed patterns. Consequently, it leads to approaches and implications rather unique and specific to the exact connectivity of interest. In functional connectivity, emphases are put on the covariance or precision matrix. For example, in Warnick et al. (2018), functional connectivity reflected in the regression error term is modeled by a Gaussian graphical model where graphs evolves according to a hidden Markov model and in the meantime are governed by a super graph. Assume that $\mathbf{y}_t$ is a vector of length N, denoting the brain signals of a subject measured on N ROIs across time $t = 1, \dots, T$ and it is the sum of some global mean $\boldsymbol{\mu}$ and K element-by-element products between design vector $X_t^k \in \mathbb{R}^N$ and the regression coefficients $\boldsymbol{\beta}_k \in \mathbb{R}^N$ corresponding to the effect of K experiment stimuli.

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{k=1}^{K} X_t^k \circ \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_t$$

$$\beta_{n,k} \sim (1 - \gamma_{n,k})\delta_0 + \gamma_{n,k}\mathcal{N}(0, \sigma_\beta^2)$$

The prior on the coefficient vector $\boldsymbol{\beta}_k$ has the spike and slab structure with a spike at 0 and Gaussian slab. The error term $\boldsymbol{\epsilon}_t$ given some state $s_t = s, s = \{1, \ldots, S\}$ is normally distributed with mean $\nvdash$ and precision matrix $\Omega_s \in \mathbb{R}^N \times \mathbb{R}^N$. The positive definite matrix Ω_s is further assumed to follow the G-Wishart distribution characterized by

$$p(\Omega \mid G, b, D) = I_G(b, D)^{-1} |\Omega|^{\frac{b-2}{2}} \exp\left(\frac{1}{2}\mathrm{tr}(\Omega D)\right),$$

where $b > 2$ is the degrees of freedom, D is a $N \times N$ positive definite matrix and I_G is the normalizing constant. The precision matrix Ω_s embeds the conditional independence relationships in a graph $G_s = (\mathcal{N}, \mathcal{E}_s)$ with N nodes $\mathcal{N}$ and state dependent edges $\mathcal{E}_s$. The state transition of s_t is modeled via a simple hidden Markov model (HMM). Edges $\mathbf{g}_{ij} = (g_{ij1}, \ldots, g_{ijs})$ with $g_{ijs} \in \{0, 1\}$ indicate the presence or absence of edge (i, j) in graph $G_s, s = 1, \ldots, S$. The edge vector is jointly modeled as

$$p(\mathbf{g}_{ij} \mid v_{ij}, \Theta) = C(v_{ij}, \Theta)^{-1} \exp(v_{ij} \mathbb{1}' \mathbf{g}_{ij} + \mathbf{g}_{ij}' \Theta \mathbf{g}_{ij}),$$

where v_{ij} controls the overall sparsity level of $\mathbf{g}_{ij}$ and $\Theta \in S \times S$ captures association between graphs at each state s. The super graph reflected in Θ encourages the selection of the same edges in related graphs. Another approach clusters the brain regions into groups that exhibit similar characteristics using classical methods such as principal components analysis (PCA) and independent components analysis (ICA) (Leonardi et al., 2013; van de Ven et al., 2004; Birn et al., 2008).

On the other hand, different tools are required for studying brain effective connectivity as it measures directed influences one brain region has on others. Nongenerative methods are not reliable when modeling the effective connectivity as they are very sensitive to signal-to-noise ratio (Zhang et al., 2015a). Main strategies to deal with effective connectivity originate from a generic state-space model incorporating differential equations that characterizes the neuronal system. An important model of this kind is the dynamic causal model (DCM) (Kiebel et al., 2007; Stephan et al., 2010; Friston et al., 2019). In general, the DCM assumes that

$$\begin{aligned}\dot{x}(t) &= f(x(t), u(t), \theta) + w(t),\\ y(t) &= g(x(t), u(t), \theta) + v(t),\end{aligned}$$

which is essentially a state-space model in continuous time. The $x(t)$ are hidden neuronal states that cannot be measured directly while the $y(t)$, as a function of $x(t)$, are observed brain signals. $u(t)$ corresponds to exogenous inputs and $w(t), v(t)$ are random fluctuations. θ is the set of model parameters that contains information on brain effective connectivity. Without knowing the specific function form of $f(x, u, \theta)$ and $g(x, u, \theta)$, the model space is huge, therefore researchers have been using a simplified version of DCM such that

$$\dot{x} = \theta^x x + u\theta^{xu} x + \theta^u u + w \tag{1.1}$$
$$\theta^x = \left.\frac{\delta f}{\delta x}\right|_{x=0} \quad \theta^{xu} = \left.\frac{\delta^2 f}{\delta x \delta u}\right|_{x=0,u=0} \quad \theta^u = \left.\frac{\delta f}{\delta u}\right|_{u=0}.$$

Here, θ^x measures the directed effect of x on its own first derivatives $\dot{x}$ while θ^u reflects the strength of influence from external u on x. θ^{xu} is the coefficient of bilinear interaction between x and u.

An alternative is the Vector Auto-regressive (VAR) model, which can be shown to be basically equivalent to DCM through mathematical manipulation. Taking the Taylor approximation of (1.1) over a time interval Δ gives the VAR model

$$\begin{aligned} x_t &= Ax_{t-\Delta} + \epsilon_t \\ A &= \exp(\Delta\theta^x) \\ \epsilon_t &= \int_0^\Delta \exp(\tau\theta^x)w(t-\tau)d\tau. \end{aligned}$$

However, the interpretation of VAR coefficients is subtle in the sense that they are no longer the magnitude of the influence that one node has on the other as in the case with DCM. Nevertheless, most literature still considers VAR coefficients as an indicator of effective connectivity. For instance, the VAR model can be further combined with prior information about human brain structure connectivity to learn about the connection between regions of interest. For instance, Chiang et al. (2017) start from the VAR model in the following form. Given the group g that subject s belongs to, the measured brain signals $\mathbf{x}_t^{(s)} \in \mathbb{R}^N, t = 1, \ldots, T$ at N regions of interest is such that

$$\mathbf{x}_t^{(s)} = \sum_{p=1}^{P} A_{g,p}^{(s)} \mathbf{x}_{t-p}^{(s)} + \boldsymbol{\epsilon}_t^{(s)} \quad , \boldsymbol{\epsilon}_t^{(s)} \sim \mathcal{N}(0, \Sigma) \quad s = 1, \ldots, S.$$

Stacking all observations $\mathbf{x}_t^{(s)}$ and error terms $\boldsymbol{\epsilon}_t^{(s)}$ across time together into vectors $X^{(s)} = (\mathbf{x}_{P+1}^{(s)\prime}, \ldots, \mathbf{x}_T^{(s)\prime})'$ and $\mathcal{E}^{(s)} = (\boldsymbol{\epsilon}_{P+1}^{(s)\prime}, \ldots, \boldsymbol{\epsilon}_T^{(s)\prime})'$, the original VAR model can be further written as

$$X^{(s)} = U^{(s)} \boldsymbol{\beta}_g^{(s)} + \mathcal{E}^{(s)}.$$

Denoting the Kronecker product by $\otimes$, they use

$$U^{(s)} = \begin{bmatrix} I_N \otimes (\mathbf{x}_P^{(s)\prime}, \ldots, \mathbf{x}_1^{(s)\prime})' \\ I_N \otimes (\mathbf{x}_{P+1}^{(s)\prime}, \ldots, \mathbf{x}_2^{(s)\prime})' \\ \vdots \\ I_N \otimes (\mathbf{x}_{T-1}^{(s)\prime}, \ldots, \mathbf{x}_{T-P}^{(s)})' \end{bmatrix},$$

and $\boldsymbol{\beta}_g^{(s)} = vec\{[A_{g,1}^{(s)}, \ldots, A_{g,P}^{(s)}]'\}$. The coefficient $\boldsymbol{\beta}_g^{(s)}$ for subject s is group specific and the prior on it is multivariate normal with mean $\Omega^{(g)}$ and covariance matrix $\Sigma^{(g)}$. The entry $\omega_k^{(g)}$ of $\Omega^{(g)}$ has a spike and slab representation where the probability of belonging to the slab incorporates prior information of structure connectivity via a probit regression.

Lastly, we address the distinction between function connectivity and effective connectivity in relation to machine learning. According to Friston (2011), functional connectivity sees higher prevalence of machine learning techniques due to its descriptive nature; scholars are interested in making diagnosis of certain diseases based on abnormal patterns observed in brain images, which is the domain where machine learning methods are more powerful. On the contrary, effective connectivity is usually used to select the more suitable model that generates the data so it is more desirable to apply generative models that shed light on the underlying mechanism.

Recent fast development of neuroscience can be partially attributed to the availability of high quality brain image data. Depending on the experiment design, brain image data are categorized into task-based and resting-state data. Both designs can be embedded in the common technologies to acquire brain images. Electroencephalography (EEG) captures electrical impulses in the brain, addressing temporal dependence over a relatively small set of locations since electrodes that can be attached to human scalp are always limited. Functional magnetic resonance imaging (fMRI), another noninvasive technology that measures blood oxygen level dependent signal (BOLD) as a proxy of brain activities. When the brain performs certain task, the active brain regions will require increased oxygen consumption and a surge of oxygen-rich blood will be observed to flow to that specific region. This change of the ratio between oxyhemogloblin and deoxyhemoglobin is recorded as the BOLD signals. The observed data usually has higher spatial resolution than temporal resolution as completing a scan usually takes 2-3 seconds, hence it reflects more spatial dependence than temporal dependence. fMRI generates results in three-dimensional voxels, which can be represented as a tensor to keep the spatial characterization.

1.3 Statistical challenges

Researchers usually use multivariate time series (EEG or fMRI signals in certain brain regions) to study brain effective connectivity. The dimension of the time series is already problematic when the research is confined to single-subject study; if one is interested in multi-subject implications, the curse of dimensionality becomes even more prominent. Furthermore, unlike brain activation studies where the parameter space is linear in time series dimension, brain effective connectivity boils down to directed interaction between brain regions, resulting in quadratically growing parameter space as the observation dimension grows. However, most of the parameters are in fact zero, this phenomenon is called small-world brain networks, characterized by dense local clustering between neighboring brain regions yet connected by few paths between pairs of brain regions from different clusters (Bassett and Bullmore, 2006; Liao et al., 2011). The sparse natural of the high dimensional parameter space brings the issue of low statistical and computational efficiency. To overcome this problem, many machine learning tools are applied to achieve sparse estimation of brain effective connectivity. The most popular one is to add regularization to coefficients via many types of penalization (Haufe et al., 2013; Hu et al., 2019). Besides statistical learning methods, the fast development of deep learning techniques such as recurrent neural networks (RNN), convolutional neural networks (CNN), graph neural networks (GNN) and generative adversarial networks (GAN) capacitate their wide application in studying brain effective connectivity (Sikka et al., 2020; Phang et al., 2019; Jun et al., 2020; Liu et al., 2020). Although these approaches can be conveniently scaled up and they have satisfactory classification and prediction performance, interpretability as well as proper quantification of uncertainty is missing.

Another statistical challenge lies in how to properly model dynamic causality across multivariate time series when assuming that interactions evolve over time. A line of research is to extend methods that are applicable to static brain effective connectivity analysis to the dynamic case. For instance, Park et al. (2018b) combine spectral dynamic causal modelling (spDCM) and parametric empirical Bayes to identify dynamic effective connectivity. As for structural equation models, the idea of dynamic effective connectivity has not received enough attention except for fairly recent work by Figueroa-Jiménez et al. (2021).

On the contrary, VAR models naturally extend to dynamic setting as long as we allow the coefficient to vary over time. Inference on the time-varying coefficients can be obtained in a Bayesian framework. This is the line of research developed in the thesis. We take a Bayesian approach and we focus and develop scalable Bayesian dynamic VAR models that can be usefully employed in the study of dynamic effective connectivity based on neuroimaging data.

1.4 Structure of the thesis

The rest of the thesis is organized as follows. In Chapter 2, we describe the static VAR model from a Bayesian perspective and we propose a new model named the Bayesian tensor VAR model with carefully designed priors that automatically determine model complexity. Chapter 2 lays the foundation for Chapter 3 as it helps us to gain a more in-depth understanding of the behavior of the proposed hierarchical prior distributions. Chapter 3 is the main part of the thesis that introduces the time-varying extension of the Bayesian tensor VAR model. We demonstrate the method's advantages over existing methods through simulation studies and real data application. Having in mind that the dynamics is essentially introduced in terms of state transitions in Chapter 3, we proceed with a different approach to modeling the VAR time-varying parameters. In Chapter 4, the parameters in TV-VAR models are governed by different stochastic processes so that the temporal evolution is smooth. Two specific process priors are employed and compared, the dynamic spike and slab prior and the dynamic shrinkage process. As a continuation of Chapter 4, we propose in Chapter 5 a novel time-varying shrinkage prior called the generalized double Pareto (GDP) process prior that admits the GDP distribution as marginal distribution. *The last two chapters are ongoing work.*

Submitted manuscripts

A paper based on Chapter 3 has been submitted and can be accessed on the arXiv at `https://arxiv.org/abs/2106.14083`

Chapter 2

Bayesian Vector Autoregressive Models

Abstract

Vector autoregressive (VAR) models are widely applied in many disciplines such as economics, neuroscience and biology. We describe the multivariate equivalent of the Yule-Walker equation in the VAR model that connects VAR coefficients with the autocovariance function. We also demonstrate VAR model's relationship with the Granger causality. For model estimation, some classical Bayesian priors for VAR models are discussed; however, they usually fail in high dimension VAR model. We propose a computationally efficient method called the Bayesian tensor VAR model with carefully devised priors and compare it with the frequentist counterpart through simulation studies, validating our method and showcasing its merit. The work can be regarded as one building block towards a better understanding of priors in time-varying VAR models in Chapter 3.

2.1 VAR model

Let $(\mathbf{y}_t)_{t\geq 1}$ be an N-dimensional time series. The vector autoregressive (VAR) model of order P assumes that $\mathbf{y}_t$ is a linear combination of the P lagged signals $\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-P}$ plus an independent noise $\boldsymbol{\epsilon}_t \in \mathbb{R}^N$,

$$\mathbf{y}_t = \begin{bmatrix} A_1, A_2, \dots, A_P \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-P} \end{bmatrix} + \boldsymbol{\epsilon}_t,$$

where $A_1, \dots, A_P$ are $N \times N$ matrices and $\boldsymbol{\epsilon}_t \overset{indep}{\sim} \mathcal{N}(0, \Sigma)$, where the symbol $\overset{indep}{\sim}$ means that they are independently distributed.

A VAR model is stable if the equation

$$A(u) \equiv |I_N - A_1 u - A_2 u^2 - \cdots - A_P u^P| = 0$$

has all roots outside the unit circle. It is weakly stationary if $E(y_t)$ does not depend on t and $Cov(y_t, y_s)$ depends only on the lag $s - t$ (assuming the involved expectations exist).

A VAR model can be represented as a dynamic linear model (DLM); more precisely, for a VAR process, it is possible to find a DLM whose measurement process (Y_t) has the same probability law as the given VAR. Several DLM representations have been proposed and the following is perhaps the most commonly used

$$\mathbf{y}_t = F\boldsymbol{\theta}_t + \mathbf{v}_t$$
$$\boldsymbol{\theta}_t = G\boldsymbol{\theta}_{t-1} + \mathbf{w}_t,$$

where F is the emission matrix, defined as $F = (1, 0, \dots, 0) \otimes \mathbf{I}_N$, where I_N is the identity matrix of dimension N and $\otimes$ denotes the Kronecker product. The state transition matrix

G is a $NP \times NP$ matrix of the form

$$G = \begin{bmatrix} A_1 & A_2 & \dots & A_{P-1} & A_P \\ \mathbf{I}_N & \mathbf{0}_N & \dots & \mathbf{0}_N & \mathbf{0}_N \\ \mathbf{0}_N & \mathbf{I}_N & \dots & \mathbf{0}_N & \mathbf{0}_N \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_N & \mathbf{0}_N & \dots & \mathbf{I}_N & \mathbf{0}_N \end{bmatrix}.$$

Consistent with this expression of G, the latent state $\boldsymbol{\theta}_t$ and the two error terms $\mathbf{v}_t$, $\mathbf{w}_t$ are

$$\boldsymbol{\theta}_t = \begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-P+1} \end{bmatrix} \quad \mathbf{v}_t \equiv 0 \quad \mathbf{w}_t = \begin{bmatrix} \boldsymbol{\epsilon}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Finally, adding the initial distribution $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0, C_0)$ completes the DLM specification. The eigenvalues of G are the reciprocal roots of $A(u)$ so that the VAR process is stable if and only if all the eigenvalues of G have modulus less than one (Prado and West, 2010). Given the autocorrelation function and partial autocorrelation function, it is possible to compute the coefficient matrix $\begin{bmatrix} A_1, A_2, \dots, A_P \end{bmatrix}$ of VAR models and vice versa. The counterpart in the univariate AR case is given by the Yule-Walker equations, whose solution can be obtained from the Durbin-Levinson recursion algorithm. Whittle (1963) proposed a multivariate extension of the Durbin-Levinson recursion algorithm applicable to VAR models. Let $R_n = E(\mathbf{y}_{t+n}\mathbf{y}_t')$ be the autocovariance function, then $R_{-n} = E(\mathbf{y}_t\mathbf{y}_{t+n}') = R_n'$. The forward and backward prediction errors are

$$\begin{aligned} \epsilon_{N,t} &= \mathbf{y}_t + A_{N,1}\mathbf{y}_{t-1} + \dots + A_{N,N}\mathbf{y}_{t-N} \\ r_{N,t} &= B_{N,N} + \dots + B_{N,1}\mathbf{y}_{t-N+1} + \mathbf{y}_{t-N}. \end{aligned} \tag{2.1}$$

Denote $R_N^\epsilon = E(\epsilon_{N,t}\epsilon'_{N,t})$, $R_N^r = E(r_{N,t}r'_{N,t})$ and

$$R_N = \begin{bmatrix} R_0 & R_1 & \dots & R_N \\ R_{-1} & R_0 & \dots & R_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_{-N} & R_{-N+1} & \dots & R_0 \end{bmatrix},$$

(2.1) implies that

$$\begin{bmatrix} \mathbf{I}_N & A_{N,1} & \dots & A_{N,N} \\ B_{N,N} & \dots & B_{N,1} & \mathbf{I}_N \end{bmatrix} R_N = \begin{bmatrix} R_N^\epsilon & \mathbf{0}_N & \dots & \mathbf{0}_N \\ B_{N,N} & \dots & \mathbf{0}_N & \mathbf{0}_N \end{bmatrix}.$$

The solution of $A_{N,1}, \dots, A_{N,N}$ first appeared in Whittle (1963) and was later developed by Morf et al. (1978) to the normalized case so that autocovariances R_n can be uniquely characterized by the partial autocorrelation function. This result is useful for simulating the coefficient matrices of stationary VAR models (Ansley and Kohn, 1986).

VAR models are widely used both for forecasting and for detecting causal relationships between variables, with the latter objective usually appealing to the notion of Granger causality. Informally, Granger causality states that if including past values of $\mathbf{y}_{i,t}$ improves the prediction of $\mathbf{y}_{j,t}$, with respect to only including past values of $\mathbf{y}_{j,t}$, then $\mathbf{y}_{i,t}$ is Granger causal for $\mathbf{y}_{j,t}$ (Granger, 1969). Consider two AR models

$$\mathbf{y}_{j,t} = \sum_{p=1}^{P} A_{p,j,j}\mathbf{y}_{j,t-p} + \epsilon^*_{j,t},$$
$$\mathbf{y}_{j,t} = \sum_{p=1}^{P} A_{p,j,j}\mathbf{y}_{j,t-p} + \sum_{p=1}^{P} A_{p,j,i}\mathbf{y}_{i,t-p} + \epsilon'_{j,t}.$$

If the variance of the prediction error $\epsilon'_{j,t}$ of $\mathbf{y}_{j,t}$ conditionally on $\mathbf{y}_{j,t-p}$ and $\mathbf{y}_{i,t-p}$, $p = 1, \dots, P$ is smaller than the variance of $\epsilon^*_{j,t}$ conditionally on $\mathbf{y}_{j,t-p}$ alone, then $\mathbf{y}_{i,t}$ Granger causes $\mathbf{y}_{j,t}$. It is therefore straightforward to infer Granger causality from the VAR coefficients. It holds that $A_{p,j,i} = 0$ for all $p = 1, \ldots, P$ if and only if $\mathbf{y}_{i,t}$ is not Granger causal for $\mathbf{y}_{j,t}$.

Partial directed coherence (PDC) is a measure of brain effective connectivity that resembles the Granger causality in frequency domain. To obtain PDC from VAR models, one first computes the Fourier transform of the coefficient matrices $A_1, A_2, \ldots, A_P$ as

$$A(\omega) = I - \sum_{p=1}^{P} A_p \exp\left(-2i\pi\omega p\right)$$

with ω being the frequency at which the Fourier transform is performed. The PDC from node i to node j at frequency ω is defined as

$$PDC_{ji}(\omega) = \frac{\mid A_{ji}(\omega) \mid}{\sqrt{\sum_{j=1}^{N} \mid A_{ji} \mid^2}},$$

where index ji denotes the jth row and ith column of a matrix. A little mathematical manipulation reveals that PDC is actually a normalized quantity as $\sum_{j=1}^{N} PDC_{ji}^2(\omega) = 1$, therefore it measures the proportion that the outflow from i to j makes up in all the outflow from i rather than the absolute strength of such information transmission. Instead of PDC at a frequency ω, people examine a frequency band; to summarize PDC over a band, a simple approach is to take the average of all PDC evaluations at each single frequency within the band of interest. For fMRI data, due to its smaller sampling frequency compared to EEG data, the frequency bands in the literature are (1) slow-5 (0.01–0.027 Hz), (2) slow-4 (0.027–0.073 Hz), (3) slow-3 (0.073–0.198 Hz), (4) slow-2 (0.198–0.5Hz) and (5) slow-1 (0.5–0.75 Hz) (Gohel and Biswal, 2015).

Due to this appealing connection to causality, VAR models are widely applied to study brain effective connectivity, which itself refers to directed influences one brain region has on other brain regions (Gorrostieta et al., 2013; Wang and Ho, 2016; Wang et al., 2016; Samdin et al., 2016; Chiang et al., 2017; Ombao et al., 2018). A common challenge in

VAR model estimation is the dimension N of the time series, as the parameter space grows quadratically with N. To obtain more reliable estimates and to avoid overfitting, one needs methodd that distinguish large coefficients from small ones. Regularization approaches such as LASSO and its variants can be applied to attain sparse coefficient estimation. From a Bayesian perspective, parsimony can be achieved by different types of shrinkage prior distributions.

2.2 Bayesian VAR model

Many classes of prior distributions have been proposed in the Bayesian literature on time series and VAR models. An early and most popular choice is the Minnesota prior, introduced by Litterman et al. (1979) and subsequently developed by other researchers at University of Minnesota. The Minnesota prior is a special case of the conjugate Iormal-inverse-Wishart prior for $A_1, \ldots, A_P$ and Σ

$$\begin{aligned} \Sigma &\sim \text{Inv-Wishart}(\Psi, \nu) \\ \alpha \mid \Sigma &\sim \mathcal{N}(a, \Sigma \otimes \Omega), \end{aligned}$$

where α is the vectorization of $\begin{bmatrix} A_1, A_2, \ldots, A_P \end{bmatrix}'$ and Ω is a $NP \times NP$ dimension matrix that conveys the prior information on the VAR coefficients. The Minnesota prior is defined though the first and second moments

$$E(A_{p,i,j} \mid \Sigma) = \begin{cases} 1 & \text{if } i = j \text{ and } p = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$Cov(A_{p,i,j}, A_{p',h,m} \mid \Sigma) = \begin{cases} \frac{\lambda^2 \Sigma_{i,h}}{p^2 \Psi_j / (\nu - N - 1)} & \text{if } j = m \text{ and } p = p' \\ 0 & \text{otherwise} \end{cases} .$$

Only the diagonal entries of A_1 has prior mean deviated from 0, which is reasonable as $\mathbf{y}_{i,t-1}$ should have the strongest power in predicting $\mathbf{y}_{i,t}$. The covariance matrix reflects prior beliefs that only $A_{p,i,j}$ and $A_{p,h,j}$ are correlated since they are both measuring the impact from $\mathbf{y}_{j,t-p}$ to $\mathbf{y}_t$. p^2 in the denominator indicates such correlation gradually dies off as the lag index p goes further into the past. Lastly, hyperparameter λ plays the role of controlling globally the correlation between VAR coefficients. For more description on the Minnesota prior and other Bayesian VAR priors such as the sum-of-coefficients prior (Doan et al., 1984) and the dummy-initial-observation prior (Sims, 1993), readers can refer to Canova (2011), Giannone et al. (2015), Koop (2013).

2.3 Bayesian tensor VAR (BTVAR) model

Although sparsity inducing priors help deliver stable coefficient estimates and better out-of-sample predictions, the parameter space is still quadratic in N, making posterior inferences problematic even for moderately large N. We propose a novel way to achieve dimension reduction in Bayesian VAR models by exploiting the lower dimensional structure in the VAR coefficient matrix such that

$$\left[A_1, A_2, \ldots, A_P\right] = \sum_{h=1}^{H} \alpha'_{3,h} \otimes \left(\alpha_{1,h} \circ \alpha_{2,h}\right),$$

where $\alpha_{1,h}, \alpha_{2,h} \in \mathbb{R}^N$ and $\alpha_{3,h} \in \mathbb{R}^P$. This formula of dimension reduction results from the decomposition of the tensor $\mathcal{A}$ that comes from stacking $A_1, \ldots, A_P$ along the lag P. By construction, $\mathcal{A}$ has order or rank $N \times N \times P$. Many tensor decomposition technique can be used to reduce the dimension (Cichocki et al., 2016); parallel factor (PARAFAC) decomposition or Canonical Polyadic (CP) decomposition is one of them. Under CP decomposition, $\mathcal{A}$ can be written as

$$\mathcal{A} = \sum_{h=1}^{H} \alpha_{1,h} \circ \alpha_{2,h} \circ \alpha_{3,h},$$

where $\circ$ denotes vector outer product. H is called the rank and it is smallest integer such that $\mathcal{A}$ can be expressed in this way. Wang et al. (2021) proposed a VAR dimension reduction method by considering the Tucker decomposition

$$\mathcal{A} = \mathcal{U} \times_1 A_1^* \times_2 A_2^* \times_3 A_3^*,$$

where $\mathcal{U}$ is another three way tensor whose ranks $r_1 \times r_2 \times r_3$ are no larger than $N \times N \times P$. $A_1^* \in \mathbb{R}^{N \times r_1}, A_2^* \in \mathbb{R}^{N \times r_2}$ and $A_3^* \in \mathbb{R}^{P \times r_3}$ are factor matrices. The operation sign $\times_1$ means mode-1 multiplication between $\mathcal{U}$ and A_1^* such that

$$(\mathcal{U} \times_1 A_1^*)_{ijk} = \sum_{r=1}^{r_1} \mathcal{U}_{rjk} (A_1^*)_{sr}$$

for $1 \leq s \leq N, 1 \leq j \leq N, 1 \leq k \leq P$. $\times_2$ and $\times_3$ are defined similarly. Combined with $l1$ penalization, their estimator is completed and called the Sparse Higher-Order Reduced-Rank (SHORR) estimator.

CP decomposition is a special case of the Tucker decomposition whose core tensor $\mathcal{U}$ has equal ranks and all entries except for the main diagonal ones are zero. We choose the CP decomposition over Tucker decomposition to reduce the dimension of VAR coefficients partially because it is convenient to specify priors on vectors. The prior that we choose is the Multiway Dirichlet generalized double Pareto (M-DGDP) prior originally proposed by Guhaniyogi et al. (2017) in tensor regression. More specifically, the prior distributions that we assume for $\alpha_{1,h}, \alpha_{2,h}$ and $\alpha_{3,h}$ are

$$\alpha_{1,h} \mid \phi_h, \tau, W_{1,h} \sim \mathcal{N}(0, \phi_h \tau W_{1,h}), \quad W_{1,h,k} \mid \lambda_{1,h,k} \sim \text{Exp}\left(\lambda^2_{1,h,k}/2\right) \quad 1 \leq k \leq N,$$

$$\alpha_{2,h} \mid \phi_h, \tau, W_{2,h} \sim \mathcal{N}(0, \phi_h \tau W_{1,h}), \quad W_{2,h,k} \mid \lambda_{2,h,k} \sim \text{Exp}\left(\lambda^2_{2,h,k}/2\right) \quad 1 \leq k \leq N,$$

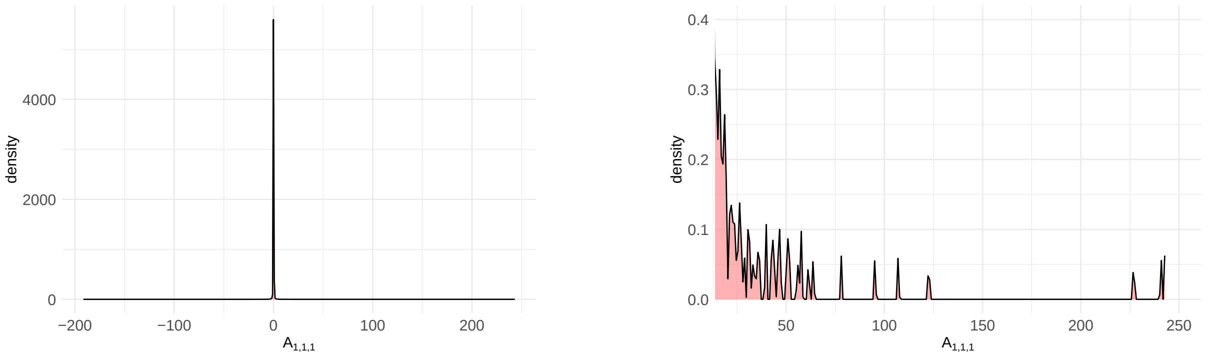
$$\phi_1, \dots, \phi_H \sim \text{Dirichlet}(\alpha, \dots, \alpha), \quad \tau \sim \text{Ga}(a_\tau, b_\tau).$$

where we further assume that the hyperparameters $\lambda_{1,h,k}, \lambda_{2,h,k} \overset{i.i.d.}{\sim} \text{Ga}(a_\lambda, b_\lambda)$. There is no natural ordering among entries of $\alpha_{1,h}$ and $\alpha_{2,h}$ so each one of them has the same prior distribution; however, for $\alpha_{3,h}$, a convincing prior belief could be to place more prior mass on entries of $\alpha_{3,h}$ corresponding to A_p's that are further into the past. Increasing shrinkage priors do this job by forcing stronger shrinkage as the index grows. A possible choice is the multiplicative Gamma processes introduced by Bhattacharya and Dunson (2011); an improvement is made by Durante (2017) to understand better the role of prior parameters. A more recent development by Legramanti et al. (2020) is based on an interpretable sequence of spike-and-slab distributions which assign increasing mass to the spike as the model complexity grows. We adapt the latter and assume that

$$\alpha_{3,h} \mid \phi_h, \tau, W_{3,h} \sim \mathcal{N}(0, \phi_h \tau W_{3,h})$$

$$W_{3,h,j} \mid z_{h,j} \sim [1 - \mathbb{1}(z_{h,j} \ \leq j)] \, \text{InvGa}(a_w, b_w) + \mathbb{1}(z_{h,j} \leq j) \, \delta_{W_\infty},$$

where each $z_{h,j}$ is a draw from a Multinomial random variable, such that $\text{pr}(z_{h,j} = l \mid w_{h,l}) = w_{h,l}$ for $l = 1, \dots, P$ with the weights $w_{h,l}$ obtained through a stick-breaking construction (Sethuraman, 1994), i.e. $w_{h,j} = v_{h,j} \prod_{l=1}^{j-1}(1 - v_{h,l})$, $v_{h,j} \sim \text{Beta}(\beta_1, \beta_2)$, $1 \leq j \leq P$. Hence, the probability of selecting the target spike is increasing with the lags j, since $P(z_{h,j} \leq j) = \sum_{l=1}^{j} w_{h,l}$. Correspondingly, the probability of choosing the Inverse Gamma slab component is $P(z_{h,j} > j) = \prod_{l=1}^{j}(1 - v_{h,l})$, i.e. decreasing with j. Higher sparsity levels for the modes $\alpha_{1,h}, \alpha_{2,h}$ and $\alpha_{3,h}$ are obtained by setting smaller



(a)

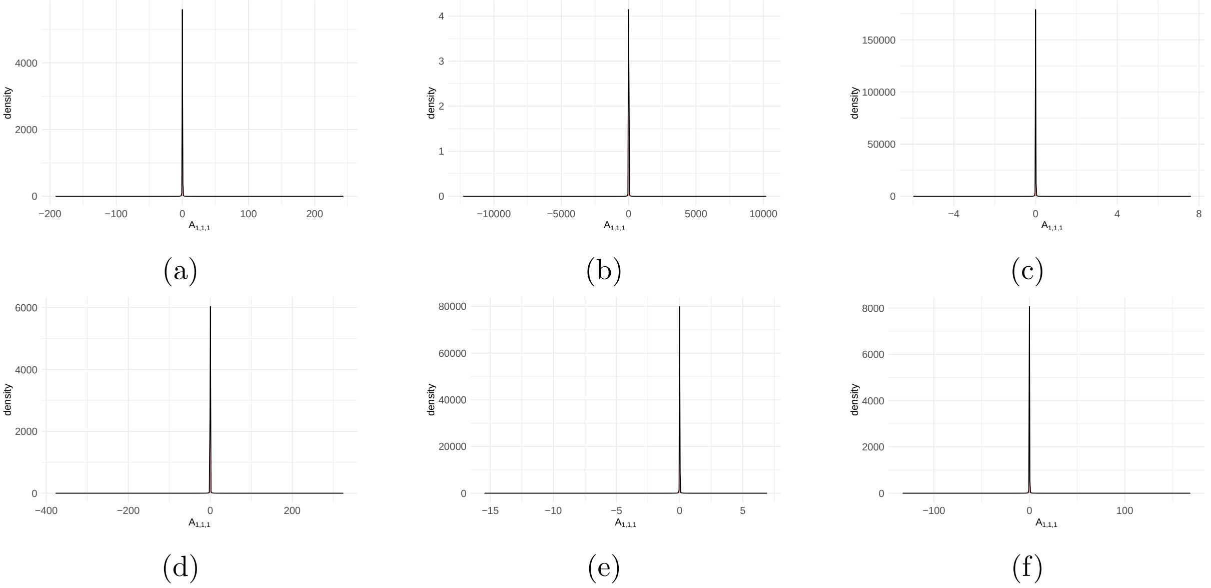
(b)

Figure 2.1: Prior distribution of $A_{1,1,1}$. $N = 4, P = 3, H = 4, a_\tau = 1/4, b_\tau = 0.3969, \alpha = 1/4, a_\lambda = 3, b_\lambda = \sqrt[6]{3}, \beta_1 = 1, \beta_2 = 1, a_w = 2, b_w = 2, W_\infty = 0.01$.

values of a_τ and b_λ relative to b_τ and a_λ, respectively. Billio et al. (2018) give the prior distribution of each cell $A_{p,i,j}$ conditionally on $\tau, \phi_1, \ldots, \phi_H$ and $W_{1,h}, W_{2,h} W_{3,h}$ in terms of Meijer G-function.

Figure 2.1 shows the prior distribution of $A_{1,1,1}$. Generally speaking, the prior is highly concentrated around 0 so that 95% of the prior probability is between -0.3174 and 0.3311. Even though the spike is prominent, the distribution has long tails as in Figure 2.1(b). Such behaviors of the prior cell distribution facilitate the distinction between signal and noise.

In Figure 2.2, we demonstrate the impact of hyperparameters on the shape of cell distributions. a_τ and b_τ are parameters of the global parameter τ, which controls the overall sparsity level of the model. Large a_τ or small b_τ results in large τ on average and weak shrinkage. The symmetric Dirichlet distribution parameter α selects the rank in CP decomposition. Small α means that most of $\phi_1, \ldots, \phi_H$ will be close to zero, favoring a lower rank decomposition. Consequently the cell prior distribution becomes more disperse since the product of normal random variables has larger variance than sum of product of normal random variables where the sum of variances of each random variables in the second case is equal to the variance of Normal random variables in the first case. This can be seen in Figure 2.2(d). Lastly, a_λ and b_λ are the hyperparameters in the generalized



(a) (b) (c)

(d) (e) (f)

Figure 2.2: Prior distribution of $A_{1,1,1}$ under different values of hyperparameter with $N = 4, P = 3, H = 4$. (a) $a_\tau = 1/4, b_\tau = 0.3969, \alpha = 1/4, a_\lambda = 3, b_\lambda = \sqrt[6]{3}, \beta_1 = 1, \beta_2 = 1, a_w = 2, b_w = 2, W_\infty = 0.01$, the rest, except for the ones specified, have the same hyperparameter values as (a). (b) $a_\tau = 4$; (c) $b_\tau = 4$; (d) $\alpha = 0.01$; (e) $a_\lambda = 10$; (f) $b_\lambda = 1$;

double Pareto (GDP) prior whose effect is very well studied in Armagan et al. (2013). The larger a_λ is, the more shrinkage toward 0 is induced on $A_{1,1,1}$ whereas b_λ has opposite implication.

For the increasing shrinkage prior on $\alpha_{3,h}$, Table 2.1 shows that as p increases, the prior distribution of $A_{p,1,1}$ becomes more and more concentrated around 0. Cell distribution in the last lag, in this case, $A_{4,1,1}$ has basically identical distribution because $W_{3,h,4}$ is always equal to 0.01 according to the definition. Hyperparameter values control the speed in which the prior shrinks towards 0. Large β_1 encourages high probability of choosing W_∞ so the variance of $A_{p,1,1}$ decays to 0 fast with p; on the contrary, if β_2 is large, the shrinkage effect increases mildly since small values of $v_{h,j}$ are more likely, which further implies that $\sum_{l=1}^{j} w_{h,l}$ grows slower to 1 by the spike and slab construction. a_w and b_w are the parameters of the slab inverse-Gamma distribution. When $z_{h,j} = 0$ and $W_{3,h,j}$ is selected from the slab, $a_w = 10$ chooses smaller variance compared to $a_w = 2$ whereas $b_w = 10$ favors more diffused prior distribution.

	0.0001	0.001	0.025	0.05	0.95	0.975	0.999	0.9999
$\beta_1 = 1$ $\beta_2 = 1$ $a_w = 2$ $b_w = 2$								
p=1	-72.9296	-15.4319	-0.5855	-0.1932	0.1921	0.5778	14.7371	65.9521
p=2	-75.4747	-12.2188	-0.3902	-0.1250	0.1251	0.3991	12.0096	55.8086
p=3	-56.3800	-8.5090	-0.2034	-0.0621	0.0589	0.1982	6.8332	39.0918
p=4	-5.3854	-1.3450	-0.0573	-0.0203	0.0201	0.0573	1.2656	6.1114
$\beta_1 = 10$ $\beta_2 = 1$ $a_w = 2$ $b_w = 2$								
p=1	-48.9256	-11.5239	-0.5056	-0.1645	0.1619	0.5171	11.8270	63.3133
p=2	-68.8159	-10.2572	-0.324	-0.1004	0.0991	0.3100	9.9001	42.3412
p=3	-40.3741	-5.0123	-0.1593	-0.0491	0.0517	0.1668	5.5007	20.6132
p=4	-6.8439	-1.3221	-0.0579	-0.0196	0.0198	0.0581	1.1933	6.0620
$\beta_1 = 1$ $\beta_2 = 10$ $a_w = 2$ $b_w = 2$								
p=1	-65.1812	-13.8639	-0.6529	-0.2208	0.2197	0.6614	14.8049	95.9737
p=2	-101.7883	-15.6717	-0.5644	-0.1835	0.1795	0.5436	13.5066	55.2943
p=3	-64.4145	-12.4110	-0.4395	-0.1419	0.1434	0.4443	11.8930	53.2769
p=4	-6.8439	-1.3221	-0.0579	-0.0196	0.0198	0.0581	1.1933	6.0620
$\beta_1 = 1$ $\beta_2 = 1$ $a_w = 10$ $b_w = 2$								
p=1	-32.9464	-5.9111	-0.2282	-0.0787	0.0777	0.2264	5.9159	22.6106
p=2	-21.0148	-5.3310	-0.1766	-0.0579	0.0589	0.1726	5.0983	20.0264
p=3	-18.5213	-2.9653	-0.1079	-0.0357	0.0396	0.1207	3.0893	13.5687
p=4	-6.459	-1.3611	-0.0572	-0.0193	0.0207	0.0610	1.4176	7.0145
$\beta_1 = 1$ $\beta_2 = 1$ $a_w = 2$ $b_w = 10$								
p=1	-163.0758	-33.6428	-1.2712	-0.4229	0.4214	1.2695	32.2362	147.4733
p=2	-168.7666	-27.3401	-0.8265	-0.2552	0.2574	0.8397	26.8986	124.8471
p=3	-126.0694	-17.4861	-0.3640	-0.1036	0.0982	0.3582	14.1988	87.4123
p=4	-5.3854	-1.3450	-0.0573	-0.0203	0.0201	0.0573	1.2656	6.1114

Table 2.1: Quantiles of the simulated prior distribution of $A_{p,1,1}$ with $N = 4, P = 4, H = 4, a_\tau = 1/4, b_\tau = 0.3969, \alpha = 1/4, a_\lambda = 3, b_\lambda = \sqrt[6]{3}, W_\infty = 0.01$.

	2	3	4	5
LASSO	0.4804 (0.2253)	0.5392 (0.2139)	0.6041 (0.2107)	0.6083 (0.2373)
SHORR	0.1128 (0.1154)	0.1311 (0.0797)	0.1759 (0.0812)	0.3150 (0.2279)
BTVAR	0.0354 (0.0253)	0.0531 (0.0438)	0.0580 (0.0444)	0.0724 (0.1053)

Table 2.2: Simulation 1: average of the squared Frobenius distance between LASSO, SHORR and BTVAR estimates and the true VAR coefficient matrix. Standard deviations are in parentheses.

2.4 Simulation studies

We compare our method with two VAR model estimators, the LASSO that was firstly applied in VAR models by Hsu et al. (2008); Basu and Michailidis (2015) and the SHORR proposed by Wang et al. (2021), through simulation studies.

In the first simulation study, we generate observations from a VAR model of moderate size with $N = 10, T = 1000, P = 5$. The underlying tensor structure of the VAR coefficient matrix has lower rank decomposition where $H = 2, 3, 4, 5$. The covariance matrix of the error term ϵ_t is diagonal and the diagonal entries all equal 10. In each setting, 100 trials are simulated and we conduct inference using the true value of P and H so there is no issue of model selection at this point. Table 2.2 reports the results of the simulation study and it can be seen that our BTVAR method outperforms the frequentist counterpart SHORR.

We investigate the performance of our method and the competing method when applied to high-dimensional VAR model in the second simulation. First simulated data are generated from a VAR model of $N = 20, T = 400, P = 8$ and the true CP factorization rank $H = 3$. When applying SHORR, we specify the lag $P = 10$ and the rank $H = 5$. The rank H will be determined using the Bayesian information criterion (BIC) and redundant lags will be recognized by the $l1$ penalization. Another 100 repetitions of data are simulated from a even higher-dimensional VAR model with $N = 40, T = 400, P = 5$ while the rank stays the same. As in the first simulation, even though we know the true

	N=20 T=400
	P=8 H=3
		N=40 N=400
		P=5 H=3

LASSO	2.5342 (0.5513)	3.1621 (0.4433)
SHORR	1.3465 (1.0503)	5.0925 (3.1179)
BTVAR	0.2098 (0.1031)	0.4386 (0.1894)

Table 2.3: Simulation 2: average of the squared Frobenius distance between LASSO, SHORR and BTVAR estimates and the true VAR coefficient matrix. Standard deviations are in parentheses.

value of H and P, we use relatively larger values to test our method's ability to automatically select the optimal model based on the data. Table 2.3 shows that our proposed method has better performance in terms of average of the squared Frobenius distance. One observation to comment is that SHORR has worse estimation of the true coefficient matrix when applied to high-dimensional VAR model with $N = 40$ even compared to the LASSO estimates. The reason for this abnormality is that the model selection index BIC fails to choose the correct rank of the tensor decomposition. In this simulation, the underlying tensor rank that generates the simulated coefficient matrix is $H = 3$ while 2 components scenarios always yield lower BIC in all 100 repetitions compared to the 3 components cases.

Chapter 3

Bayesian Time-Varying Tensor Vector Autoregressive Models for Dynamic Effective Connectivity

Abstract

Recent developments in functional magnetic resonance imaging (fMRI) investigate how some brain regions directly influence the activity of other regions of the brain *dynamically* throughout the course of an experiment, i.e. the so-called dynamic effective connectivity. Time-varying vector autoregressive (TV-VAR) models have been employed to draw inferences for this purpose, but they are very computationally intensive, since the number of parameters to be estimated increases quadratically with the number of time-series. In this paper, we propose a computationally efficient Bayesian time-varying VAR approach for modeling high-dimensional time series. The proposed framework employs a tensor decomposition for the VAR coefficient matrices at different lags. Dynamically varying connectivity patterns are captured by assuming that at any given time only a subset of components in the tensor decomposition is active. Latent binary time-series select the active components at each time via a convenient Ising prior specification. The proposed prior structure encourages sparsity in the tensor structure and allows to ascertain model complexity through the posterior distribution. More specifically, sparsity-inducing priors are employed to allow for global-local shrinkage of the coefficients, to determine automatically the rank of the tensor decomposition and to guide the selection of the lags of the auto-regression. We show the performances of our model formulation via simulation studies and data from a real fMRI study involving a book reading experiment.

Keywords: Time-varying vector autoregressive models, Tensor factorization, Brain effective connectivity

3.1 Introduction

A primary goal of many functional magnetic resonance imaging (fMRI) experiments is to investigate the integration among different areas of the brain in order to explain how cognitive information is distributed and processed. Neuroscientists typically distinguish between *functional connectivity*, which measures the undirected associations, or temporal correlation, between the fMRI time series observed at different locations, and *effective connectivity*, which estimates the directed influences that one brain region exerts onto other regions (Friston, 2011; Zhang et al., 2015b; Durante and Guindani, 2020). In this manuscript, we focus on modeling effective connectivity via a vector auto-regression (VAR) model, a widely-employed framework for estimating temporal (Granger) casual dependence in fMRI experiments (see, e.g. Gorrostieta et al., 2013; Chiang et al., 2017). In addition, in our motivating dataset, it is envisaged that the connectivity patterns may vary *dynamically* throughout the course of the fMRI experiment. Recent literature in the neurosciences has recognized the need to describe changes in brain connectivity in response to a series of stimuli in task-based experimental settings or because of inherent spontaneous fluctuations in resting state fMRI (Hutchison et al., 2013; Taghia et al., 2017; Park et al., 2018a; Warnick et al., 2018; Zarghami and Friston, 2020a). Samdin et al. (2016) and Ombao et al. (2018) have recently employed a Markov-switching VAR model formulation to characterize dynamic connectivity regimes among a few selected EEG channels. More recently, Li et al. (2020) have developed a stochastic block-model state-space multivariate auto-regression for investigating how abnormal neuronal activities start from a seizure onset zone and propagate to otherwise healthy regions using intracranial EEG data.

VAR models are computationally intensive for analyzing high-dimensional time-series, since the number of parameters to be estimated increases quadratically with the number of time-series, easily surpassing the number of observed time points. Hence, several approaches have been proposed to enforce sparsity of the VAR coefficient matrix, either by using penalized-likelihood methods (Shojaie and Michailidis, 2010; Basu and Michailidis, 2015) or - in a Bayesian setting - by using several types of shrinkage priors (Primiceri, 2005; Koop, 2013; Giannone et al., 2015). Alternatively, dimension reduction techniques have been employed to reveal and exploit a lower dimensional structure embedded in the parameter space. For example, Velu et al. (1986) decompose the VAR coefficient matrix as the product of lower-rank matrices. More recently, Billio et al. (2018) consider a tensor decomposition to model the (static) parameters of a time-series regression. Wang et al. (2021) have proposed an L_1- penalized-likelihood approach where a tensor decomposition is used to express the elements of the VAR coefficient matrices.

In this paper, motivated by an experimental study on dynamic effective connectivity patterns arising when reading complex texts, we propose a computationally efficient time-varying Bayesian VAR approach for modeling high-dimensional time series. Similarly as in Wang et al. (2021), we assume a tensor decomposition for the VAR coefficient matrices at different lags. A novel feature of the proposed approach is that we capture dynamically varying connectivity patterns by assuming that – at any given time – the VAR coefficient matrices are obtained as a mixture of just a subset of active components in the tensor decomposition. This mixture representation relies on latent indicators of brain activity, that we model through an innovative use of an Ising prior on the time-domain, to select what components are active at each time. With respect to Hidden Markov Models – typically employed in the fMRI literature to capture transitions across brain states dynamically over time – the Ising prior models the time-varying activations as a function of only two parameters. The resulting binary time series still maintains a Markovian dependence, but the Ising prior naturally assigns a higher probability mass to non-active (zeroed) components to encourage sparsity of representation and it favors similar selections at two consecutive time points, reflecting the prior belief that the coefficients are changing slowly over time. Furthermore, we show that the Ising prior can be represented as the joint distribution of a so-called (new) discrete autoregressive moving average (NDARMA) model (Jacobs and Lewis, 1983), a result which is helpful for prior elicitation.

The remaining components of the model are designed to encourage sparsity in the tensor structure and to ascertain model complexity directly from the data through the posterior distribution. In particular, we employ a multi-way Dirichlet generalized double Pareto prior (Guhaniyogi et al., 2017) to allow for global-local shrinkage of the VAR coefficients and to determine automatically the effective rank of the tensor decomposition. A further feature of our approach is that we assume an *increasing*-shrinkage prior (Legramanti et al., 2020) to guide the selection of the lags of the auto-regression, without the need for ranking different models based on model selection information criteria.

The rest of the paper is organized as follows. In Section 3.2, we formulate the time-varying tensor model and elucidate how to obtain dimension reduction via a tensor decomposition into a set of latent base matrices and binary indicators of connectivity patterns over time. In Section 3.3, we describe the Ising prior specification on the temporal transitions, as well as the sparsity-inducing priors on the active elements of the tensor decomposition. In Section 3.3.4 we discuss posterior computation and inference. Results of the simulation studies as well as the real data application are shown in Section 3.4 and Section 3.5 respectively. Lastly, Section 3.6 provides some concluding remarks and future work.

3.2 Time-Varying Tensor VAR (TVT-VAR) model for Effective Connectivity

In this Section, we introduce the proposed time-varying tensor VAR (TVT-VAR) specification for studying dynamic brain effective connectivity. Let $\mathbf{y}_t$ be an N-dimensional vector for $t = 1, \dots, T$. Each time-series data $(y_{i1}, \dots, y_{iT})$ represents the fMRI BOLD signal recorded at voxel or region of interest (ROI) i, $i = 1, \dots, N$. The TVT-VAR model of order P assumes that $\mathbf{y}_t$ is a linear combination of the P lagged signals $\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-P}$ plus an independent noise $\boldsymbol{\epsilon}_t \in \mathbb{R}^N$,

$$\mathbf{y}_t = \begin{bmatrix} A_{1,t}, A_{2,t}, \ldots, A_{P,t} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-P} \end{bmatrix} + \boldsymbol{\epsilon}_t, \tag{3.1}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \Sigma)$ and the linear coefficients $A_{j,t},\ j = 1, \ldots, P$ are $N \times N$ matrices, assumed to vary across t, $t = 1, \ldots, T$. We assume that Σ is time-invariant and diagonal, and we focus on the coefficient matrices $\begin{bmatrix} A_{1,t}, A_{2,t}, \ldots, A_{P,t} \end{bmatrix}$. If needed, the assumption on Σ can be appropriately relaxed. The number of coefficients to be estimated is $(T - P) \times N^2 \times P + N$; hence, it is not possible to use the conventional ordinary least square estimator. We propose to address the issue following multiple simultaneous strategies.

First, we model the dynamic coefficient matrix as a time-varying mixture of H latent static base coefficient matrices. More specifically, let $(\gamma_{h,t})_{t \geq P+1}$ be a binary-valued time series, $h = 1, \ldots, H$. Then, we assume

$$\begin{bmatrix} A_{1,t}, A_{2,t}, \ldots, A_{P,t} \end{bmatrix} = \sum_{h=1}^{H} \gamma_{h,t} \begin{bmatrix} A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h} \end{bmatrix} \tag{3.2}$$

that is, for any t, each VAR coefficient matrix $A_{j,t}$ is a composition of the subset of those base matrices $A^*_{j,h}$ for which $\gamma_{h,t} = 1$, $h = 1, \ldots, H$, $j = 1, \ldots, P$. The binary $\gamma_{h,t}$'s can be interpreted as indicators of latent individual or experimental conditions. For example, Gorrostieta et al. (2013) have previously proposed the use of a known binary indicator for comparing connectivity across experimental conditions (e.g., active vs. rest in task-based fMRI). Instead, we infer the latent $\gamma_{h,t}$ from the data to explore latent varying patterns in brain effective connectivity that are not necessarily tied to experiment conditions. Similarly, $A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}$ can be interpreted as *latent base matrices.* Compared with estimating $N^2 \times P$ time-series of length $(T - P)$ in the initial model specification, our formulation (3.2) requires estimating $N^2 \times P$ base matrices $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$ and the dynamics of the VAR coefficient matrices are now governed by the temporal dependence between the $\gamma_{h,t}$'s, $h = 1, \ldots, H$, $t = P+1, \ldots, T$. A natural choice is to set the $\gamma_{h,t}$'s as independent across different mixing components h, but envision some Markovian dependence over different time points t (see Section 3.3.1).

Despite the reduced dimensionality, espression (3.2) remains highly parameterized. Hence, we further propose to stack each set of matrices $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$ into a three-way tensor $\mathcal{A}^*_h$ of size $N \times N \times P$ and then apply a PARAFAC decomposition to achieve an increased reduction in the number of estimands. Along with the Tucker decomposition, the PARAFAC decomposition is often employed for tensor dimension reduction due to its straightforward interpretation and implementation. Hoff (2015) has proposed the use of a Tucker product for dimension reduction in a general multi-linear tensor regression framework for the analysis of longitudinal relational data. Tensors have been used before also in the neuroimaging literature for detecting activations via tensor regression approaches (Zhou et al., 2013; Guhaniyogi et al., 2017), but – to our knowledge – their use for studying effective connectivity within VAR models has not been yet explored. In general, a $q_1 \times q_2 \times \cdots \times q_M$ tensor $\mathcal{A}$ is said to admit a rank-R PARAFAC decomposition if R is the smallest integer such that $\mathcal{A}$ can be written as

$$\mathcal{A} = \sum_{r=1}^{R} \alpha_{1,r} \circ \alpha_{2,r} \circ \cdots \circ \alpha_{M,r}, \tag{3.3}$$

where $\circ$ indicates the vector outer product and $\alpha_{m,r} \in \mathbb{R}^{q_m}, m = 1, \ldots, M$ are the *tensor margins* of each mode. In Figure 3.1, we show a simple graphical illustration of the PARAFAC decomposition of a three-way tensor. The tensor representation is important to reduce dimension but the inferential interest is on recovering the temporal patterns of the VAR coefficients. At this regard, it is important to note that, while inference on (3.3) may suffer from identifiability issues, the $A_{j,t}$ remain identifiable. Indeed, the

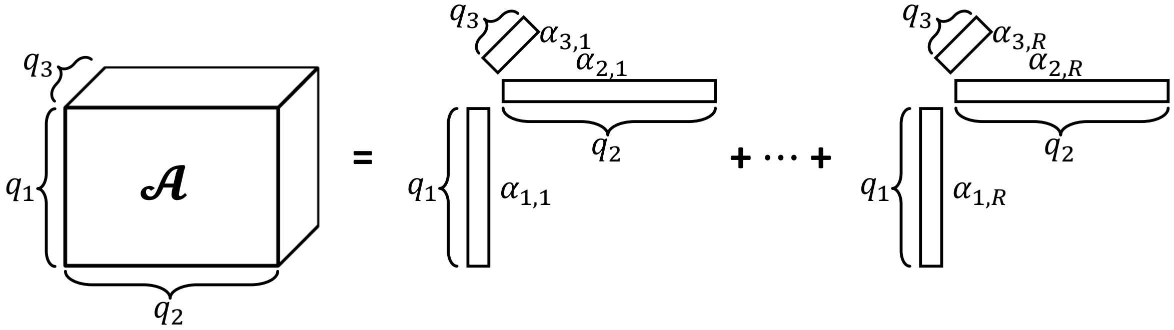


Figure 3.1: An illustrative example of a Rank-R PARAFAC decomposition of a three-way $q_1 \times q_2 \times q_3$ tensor $\mathcal{A}$.

decomposition of $\mathcal{A}$ is invariant under any permutation of the component indices r so that $\mathcal{A} = \sum_{r=1}^{R} \alpha_{1,\Pi(r)} \circ \alpha_{2,\Pi(r)} \circ \cdots \circ \alpha_{M,\Pi(r)}$ for any permutation $\Pi(\cdot)$ of the index set $\{1, 2, \ldots, R\}$. Moreover, $\mathcal{A}$ is not altered by rescaling, i.e. $\mathcal{A} = \sum_{r=1}^{R} \alpha^*_{1,r} \circ \alpha^*_{2,r} \circ \cdots \circ \alpha^*_{M,r}$ where $\alpha^*_{m,r} = \nu_{m,r}\alpha_{m,r}$ for any set of multiplying factors $\nu_{m,r}$ such that $\prod_{m=1}^{M} \nu_{m,r} = 1$. However, the temporal pattern of the VAR coefficients, i.e. the product of the margins, remains identifiable.

In our model, we assume that the base tensors $\mathcal{A}^*_h$ admit a rank-1 PARAFAC decomposition to allow for a more parsimonious parameterization, i.e.

$$\mathcal{A}^*_h = \alpha_{1,h} \circ \alpha_{2,h} \circ \alpha_{3,h} \quad h = 1, \ldots, H,$$

where $\alpha_{1,h}, \alpha_{2,h} \in \mathbb{R}^N$. The tensor margin $\alpha_{3,h} \in \mathbb{R}^P$ denotes the lag mode, i.e. the tensor margin related to the order of the VAR models. Since the influence of past variables is expected to diminish with increasing lags, the entries of $\alpha_{3,h}$ should also be expected to decreasing with increasing lags. In Section 3.3.2, we describe prior specifications that enforce sparsity in $\alpha_{1,h}$ and $\alpha_{2,h}$ and increasing shrinkage in $\alpha_{3,h}$.

We can further express the original matrix $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$ by rearranging the modes of the tensor decomposition as follows,

$$\left[A_{1,h}^{*}, A_{2,h}^{*}, \ldots, A_{P,h}^{*}\right] = \alpha_{3,h}' \otimes (\alpha_{1,h} \circ \alpha_{2,h}) \quad h = 1, \ldots, H;$$

where $\otimes$ denotes the Kronecker product. The construction is useful to highlight a sequence of constraints on the coefficient matrix: the element-by-element ratio between $A_{1,h}^{*}$ and $A_{2,h}^{*}$ is proportional to the ratio between the first two entries of $\alpha_{3,h}$, and similarly for subsequent lags. In summary, after matricization, the set of TVT-VAR coefficients can be expressed as a mixture of only an active subsets of components as

$$\left[A_{1,t}, A_{2,t}, \ldots, A_{P,t}\right] = \sum_{h=1}^{H} \gamma_{h,t}\, \alpha_{3,h}' \otimes (\alpha_{1,h} \circ \alpha_{2,h}),$$

where latent binary time-series $\gamma_{h,t}$'s select the active components at each time, $h = 1, \ldots, H$. Alternatively, we can stack the TVT-VAR coefficients in (3.2) and obtain the tensor $\mathcal{A}_t$, which can be written as

$$\mathcal{A}_t = \sum_{h=1}^{H} \gamma_{h,t}\, \mathcal{A}_h^{*} = \sum_{h=1}^{H} \gamma_{h,t}\, (\alpha_{1,h} \circ \alpha_{2,h} \circ \alpha_{3,h}).$$

The expressions above highlight that the proposed tensor decomposition reduces the number of parameters to $H(T-P)+H(2N+P)$, i.e. linear in the observation size N, instead of N^2 without the tensor reparameterization.

Finally, we should note that also Sun and Li (2019) have recently proposed to stack a series of dynamic tensors to form a higher order tensor. In their approach, the data are observed tensors to be clustered over time along the modes generated via the PARAFAC decomposition. Instead, in our approach the data are multivariate time series and the tensor structure is used to construct a lower dimensional parameter space for the unknown VAR coefficients to be estimated. In addition, Sun and Li (2019) achieve smoothness in the parameters through a fusion structure that penalizes discrepancies between neighboring entries in the same tensor margin. Instead, we follow a Bayesian approach and further encourage a contiguous structure by means of the Ising prior distribution detailed in the following section.

3.3 Prior specifications

3.3.1 Ising prior on temporal transitions

The sequence of latent indicators $\gamma_{h,t}$ determines the time-varying activations of the latent base matrices in the VAR model (3.1)–(3.2). Hidden Markov Models based on homogeneous temporal transitions have been used in recent neuroimaging literature to describe temporal variations of functional connectivity patterns (Baker et al., 2014; Vidaurre et al., 2017; Warnick et al., 2018). Here, we propose an Ising prior specification on the time domain that retains the Markovian dependence but does allow to model the time-varying activations as a function of only two parameters for each of the bases, one parameter capturing general sparsity and the other capturing the strength of dependence between adjacent time points. More specifically, we assume that, independently for each h, the binary state process $(\gamma_{h,t})_{t>P}$ is characterized by joint probability mass functions

$$\begin{aligned} &P\left(\gamma_{h,P+1},\ldots,\gamma_{h,T} \mid \theta_h,\kappa_h\right) \\ &\qquad \propto \exp\left(\theta_h\gamma_{h,P+1}+\sum_{t=P+2}^{T-1}\theta_h^*\gamma_{h,t}+\theta_h\gamma_{h,T}+\sum_{t=P+1}^{T-1}\kappa_h\gamma_{h,t}\gamma_{h,t+1}\right). \end{aligned} \tag{3.4}$$

Equation (3.4) defines an Ising model, i.e. an undirected graphical model or Markov random field involving the binary random vector $\boldsymbol{\gamma}_h = (\gamma_{h,P+1},\ldots,\gamma_{h,T}) \in \{0,1\}^{T-P}$, $h=1,\ldots,H$ (see, e.g., Wainwright and Jordan, 2008). The parameters θ_h and θ_h^* can be interpreted as *sparsity* parameters, since they correspond to the probability of activation for component h at each time t, irrespective of the status at $t-1$ and $t+1$. Positive values of θ_h and θ_h^* increase the probability that $\gamma_{h,t}=1$; instead, negative values of

θ_h and θ_h^* increase the probability that $\gamma_{h,t} = 0$, $t = P+1, \ldots, T$. The parameter κ_h captures the effect of the interaction between $\gamma_{h,t}$ and $\gamma_{h,t+1}$. In particular, when $\kappa_h > 0$, the probability that $\gamma_{h,t}$ and $\gamma_{h,t+1}$ are both non-zero is larger.

The Ising prior (3.4) can be seen as a specific instance of a multivariate Bernoulli distribution, as defined by Dai et al. (2013). In particular, in the following we show how the proposed prior is equivalent to a binary discrete autoregressive NDARMA(1) model (Jacobs and Lewis, 1983; MacDonald and Zucchini, 1997; Jentsch and Reichmann, 2019). For notational simplicity, we focus on a single time series $\gamma_{h,t}$, and we omit the subscript h for the remainder of the Section. We start by recalling that a NDARMA(1) process is a binary time series that satisfies

$$\gamma_t = a_t \, \gamma_{t-1} + (1 - a_t) \, \epsilon_t, \quad t = 1, \ldots, T \tag{3.5}$$

where $a_t \overset{i.i.d.}{\sim} \text{Bern}(p_1)$, and $\epsilon_t \overset{i.i.d.}{\sim} \text{Bern}(p_2)$, with a_t and ϵ_t independent. The initial condition assumes $\gamma_1 \sim \text{Bern}(p_2)$. The NDARMA(1) model has a Markovian dependence structure, with transition probabilities

$$P(\gamma_t \mid \gamma_{t-1}) = p_1 \, \mathbb{1}(\gamma_t = \gamma_{t-1}) + (1 - p_1) \, p_2^{\gamma_t} \, (1 - p_2)^{\gamma_{t-1}},$$

for $\gamma_t, \gamma_{t-1} \in \{0, 1\}$. Moreover, marginally $\gamma_t \sim \text{Bern}(p_2)$. Intuitively, the autocorrelation function at lag 1 of the NDARMA time series is always positive, meaning that γ_t and γ_{t+1} tend to assume the same value, consistent with the contiguous behavior that we would like the Ising prior (3.4) to encourage by setting $\kappa > 0$. Then, in a NDARMA(1) model, the joint probability mass function of $\gamma_1, \ldots, \gamma_T$ can be obtained as

$$\begin{aligned} p_{\gamma_1,\ldots,\gamma_T} &= P(\gamma_1, \ldots, \gamma_T) = P(\gamma_1) \prod_{t=2}^{T} P(\gamma_t|\gamma_{t-1}) \\ &= p_2^{\gamma_1} (1-p_2)^{\gamma_1} \prod_{t=2}^{T} \left\{ p_1 \, \mathbb{1}(\gamma_t = \gamma_{t-1}) + (1-p_1)\, p_2^{\gamma_t} (1-p_2)^{\gamma_{t-1}} \right\}. \end{aligned}$$

For instance, the probability of a zero-sequence, $p_{0\ldots0} = P(\gamma_1 = 0, \ldots, \gamma_T = 0)$, equals $(1-p_2) \prod_{t=2}^{T} (p_1 + (1-p_1)(1-p_2))$. Let $n = \sum_{t=1}^{T} \gamma_t$ indicate the total number of active indicators γ_t's along the entire time-series, and let $\{j_1, \ldots, j_n\} \subset \{1, \ldots, T\}$ denote the subset of times j_r where $\gamma_{j_r} = 1$, $r = 1, \ldots, n$. Then, in a NDARMA(1) model, the vector $(\gamma_1, \ldots, \gamma_T)$ follows a multivariate Bernoulli distribution, as defined in Dai et al. (2013). More specifically, the joint distribution can be rewritten as

$$P(\gamma_1, \ldots, \gamma_T) = p_{0\ldots0}^{\prod_{t=1}^{T}(1-\gamma_t)} \, p_{10\ldots0}^{\gamma_1 \prod_{t=2}^{T}(1-\gamma_t)} \, p_{01\ldots0}^{(1-\gamma_1)\,\gamma_2 \prod_{t=3}^{T}(1-\gamma_t)} \cdots p_{1\ldots1}^{\prod_{t=1}^{T} \gamma_t}.$$

Let $B^{j_1 j_2 \cdots j_r}(\boldsymbol{\gamma}) = \gamma_{j_1} \gamma_{j_2} \cdots \gamma_{j_r}$ define a general interaction function among a subset $\{j_1, \ldots, j_r\}$ of the γ_t's. Dai et al. (2013) show that the multivariate Bernoulli distribution is a member of the exponential family, i.e. the joint probability of $(\gamma_1, \ldots, \gamma_T)$ can be reweritten as

$$P(\gamma_1, \ldots, \gamma_T) \propto \exp\left(\sum_{n}^{T} \left(\sum_{1 \le j_1 < j_2 < \cdots < j_n \le T} f_T^{j_1 j_2 \ldots j_n} \, B^{j_1 j_2 \cdots j_r}(\boldsymbol{\gamma}) \right) \right), \tag{3.6}$$

where $f_T^{j_1 j_2 \ldots j_n}$ is the natural parameter defined by the equation

$$\exp\left(f_T^{j_1 j_2 \ldots j_n}\right) = \prod_{\{\text{even } \# \text{ of 0's in } \gamma_{j_1}, \ldots, \gamma_{j_n}\}} p^*_{T, j_1 \ldots j_n} \Bigg/ \prod_{\{\text{odd } \# \text{ of 0's in } \gamma_{j_1}, \ldots, \gamma_{j_n}\}} p^*_{T, j_1 \ldots j_n},$$

with $p^*_{T, j_1 \ldots, j_n}$ denoting the probability that the γ_t's at times $j_1, \ldots, j_n$ are $\gamma_{j_1}, \ldots, \gamma_{j_n}$ and all others are zero. It is easy to see that the Ising prior (3.4) is a special case of the equation

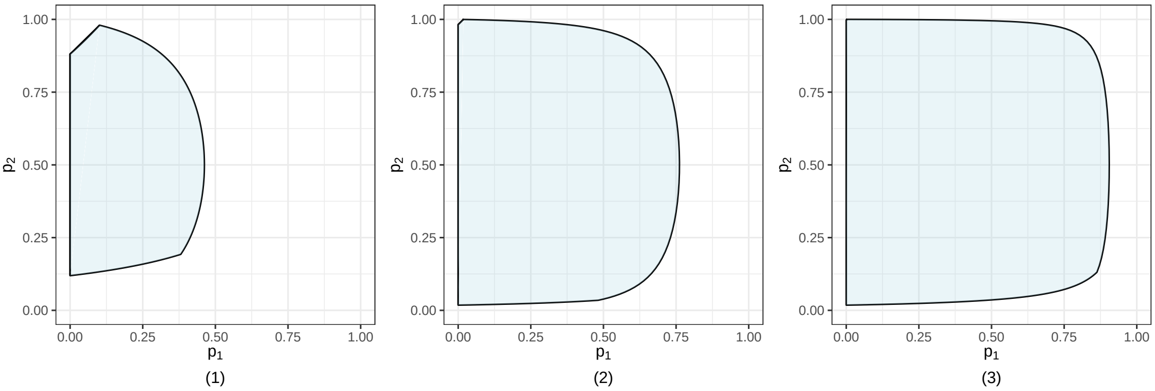


Figure 3.2: An illustration of the mapping between the parameters values of (p_1, p_2) of the NDARMA(1) model (3.5) and the parameters (θ, κ) of the Ising prior (3.3.1). The domain of (θ, κ) constrains the admissible range of (p_1, p_2). The three panels illustrate the domain of (p_1, p_2) corresponding to increasing domains of (θ, κ): (1) $-2 < \theta < 2, 0 < \kappa < 2$; (2) $-4 < \theta < 4, 0 < \kappa < 4$; (3) $-4 < \theta < 6, 0 < \kappa < 6$.

(3.6) by setting $\theta = f_T^{P+1} = f_T^T$, $\theta^* = f_T^{P+2} = \cdots = f_T^{T-1}$, $\kappa = f_T^{P+1,P+2} = \cdots = f_T^{T-1,T-2}$. Since NDARMA(1) models encode a Markov dependence of order 1, the coefficient $f_T^{j_1 \ldots j_n}$ associated with $\gamma_{j_1} \cdots \gamma_{j_n}$ is zero for $n \geq 3$.

The following proposition maps the parameters (θ, κ) in the Ising prior (3.4) to the parameters (p_1, p_2) in the NDARMA(1) model 3.5:

Proposition 1. *The probability law of the NDARMA*(1) *model in* (3.5) *can be expressed as in* (3.4)*. In particular, the parameters* (θ, κ) *are obtained as a function of the parameters* p_1*,* p_2 *in* (3.5) *as*

$$e^\theta = \frac{p_2(1-p_1)}{p_1 + (1-p_2)(1-p_1)}, \quad e^\kappa = \frac{p_1 + p_2(1-p_2)(1-p_1)^2}{p_2(1-p_2)(1-p_1)^2},$$

$$\exp(\theta^*) = \frac{p_2(1-p_2)(1-p_1)^2}{(p_1 + (1-p_2)(1-p_1))^2} = \frac{e^\theta(e^\theta+1)}{e^{\theta+\kappa}+1}.$$

Inversely,

$$p_1 = \frac{e^\theta(e^\kappa - 1)}{(e^{\theta+\kappa}+1)(e^\theta+1)}, \quad p_2 = \frac{e^\theta(e^{\theta+\kappa}+1)}{e^{2\theta+\kappa}+2e^\theta+1}.$$

The previous result is helpful in setting the prior distributions for the parameters in (3.4), as it highlights hidden constraints among the parameters and how the domain of (θ, κ) constrains the admissible range of (p_1, p_2). Indeed, we notice that the transformation is bijective, since θ^* can be expressed as a function of the pair (θ, κ). The parameters κ and p_1 have the same sign, since they both indicate the strength of the dependence between two neighboring γ_{t-1} and γ_t. It also follows that, due to the positiveness of κ, θ^* is always smaller than θ. As an illustration of the complex dependencies induced by the mapping between $(\theta, \theta^*, \kappa)$ in (3.4) and (p_1, p_2) in (3.5), Figure 3.2 compares the range of (p_1, p_2) for different intervals of values of (θ, κ). For example, if $0 < \kappa < 2$ and $-2 < \theta < 2$ (panel 1), the corresponding set of NDARMA(1) models is limited to a subset of those with $p_1 < 0.5$. As the domain of (θ, κ) expands, the set of induced NDARMA(1) models also expands. Shrinking $\gamma_{h,t}$ towards zero is desirable for regularization purposes, which corresponds to allowing negative values of the parameters θ_h. At the same time, too much shrinkage may hamper our ability to identify latent base patterns that are recurrent, as the shrinkage may result in too low estimates of p_1 and p_2. Indeed, it is well known that the prior specification of the parameters of a Ising model needs to be conducted with care, in order to avoid the phenomenon of phase-transition (Li and Zhang, 2010; Li et al., 2015). In statistical physics, a phase-transition refers to a sudden change from a disordered (non-magnetic) to an ordered (magnetic) state at low temperatures. In Bayesian variable selection, the phase-transition has been associated to values of the parameter space that lead to selecting either all or none of the tested variables. These considerations motivate our suggestion of a proper uniform distribution on the parameters (θ, κ) over a closed interval in $\mathbb{R}^2$ for posterior inference. More specifically, we assume that θ_h lies between $[\theta_{h,\min}, \theta_{h,\max}]$ with lower limit $\theta_{h,\min} < 0$ and upper limit $\theta_{h,\max} > 0$. We have found that choosing $\theta_{h,\min} = -4$ and $\theta_{h,\max} = 4$ ensures a proper exploration of the parameter space and appears to avoid phase transitions. Similarly, for κ_h, we encourage a contiguous structure where $\gamma_{h,t}$ and $\gamma_{h,t+1}$ are simultaneously selected by assuming that κ_h is positive with a uniform prior on $\kappa_h \in [0, \kappa_{h,\max}]$ with $\kappa_{h,\max} > 0$. Also in this case an upper limit $\kappa_{h,\max} = 4$ appears to ensure both reasonably good inference on the time-varying coefficients and computational efficiency.

3.3.2 Sparsity-inducing priors

In addition to the dimension reduction achieved through the PARAFAC decomposition, we seek further shrinkage of the tensor margins' parameters. For that purpose, we consider priors that shrink the parameters toward zero, enabling a sparse representation of the VAR coefficents and more interpretable estimation of the connectivity patterns. In particular, for the elements of the tensor margins $\alpha_{1,h}$ and $\alpha_{2,h}$ we consider a multi-way Dirichlet generalized double Pareto prior (Guhaniyogi et al., 2017), whereas for the lag margin $\alpha_{3,h}$, we consider an increasing shrinkage prior, so that higher-order lags are penalized. More specifically, we assume that $\alpha_{1,h}$ and $\alpha_{2,h}$, which determine the rows and columns of the original VAR coefficient matrix, are normally distributed with zero mean and variance-covariance matrix $\tau\phi_h W_{1,h}$, with $W_{1,h}$ diagonal. The parameter τ is a global scale parameters that follows a Gamma distribution, whereas the $\phi_1, \ldots, \phi_H$ are local scale parameters that follow a symmetric Dirichlet distribution. The diagonal elements of the covariance have a generalized double Pareto prior:

$$\begin{gathered}
\alpha_{1,h} \mid \phi_h, \tau, W_{1,h} \sim \mathcal{N}(0, \phi_h \tau W_{1,h}), \quad W_{1,h,k} \mid \lambda_{1,h,k} \sim \mathrm{Exp}\left(\lambda^2_{1,h,k}/2\right) \quad 1 \leq k \leq N, \\
\alpha_{2,h} \mid \phi_h, \tau, W_{2,h} \sim \mathcal{N}(0, \phi_h \tau W_{1,h}), \quad W_{2,h,k} \mid \lambda_{2,h,k} \sim \mathrm{Exp}\left(\lambda^2_{2,h,k}/2\right) \quad 1 \leq k \leq N, \\
\phi_1, \ldots, \phi_H \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \quad \tau \sim \mathrm{Ga}(a_\tau, b_\tau).
\end{gathered}$$

where we further assume that the hyperparameters $\lambda_{1,h,k}, \lambda_{2,h,k} \overset{i.i.d.}{\sim} \mathrm{Ga}(a_\lambda, b_\lambda)$.

By setting $a_\tau = H\alpha$, one can obtain tractable full conditionals distributions for τ and $(\phi_1, \ldots, \phi_H)$, since the full conditional for τ is a generalized inverse Gaussian distribution and the full conditionals for $(\phi_1, \ldots, \phi_H)$ are normalized generalized inverse Gaussian random variables (Guhaniyogi et al., 2017).

For the lag mode parameter, $\alpha_{3,h}$, we maintain a normal prior with a global-local structure on the covariance, i.e. $\alpha_{3,h} \mid \phi_h, \tau, W_{3,h} \sim \mathcal{N}(0, \phi_h \tau W_{3,h})$. However, we provide for a cumulative shrinkage effect on the diagonal entries of $w_{3,h}$ to encourage the estimation of a small number of lags. More specifically, we employ a cumulative shrinkage prior (Legramanti et al., 2020), i.e. a prior which induces increasing shrinkage via a sequence of spike and slab distributions assigning growing mass to a target spike value as the model complexity grows. Let W_∞ indicate the target spike, e.g. $W_\infty = 0$ or $W_\infty = 0.05$ to avoid degeneracy of the normal distribution to a point mass and improve computational efficiency (George and McCulloch, 1993). Then, for any j, $1 \leq j \leq P$, we assume

$$W_{3,h,j} \mid z_{h,j} \sim [1 - \mathbb{1}(z_{h,j} \ \leq j)] \, \mathrm{InvGa}(a_w, b_w) + \mathbb{1}(z_{h,j} \leq j) \, \delta_{W_\infty},$$

where each $z_{h,j}$ is a draw from a Multinomial random variable, such that $\mathrm{pr}(z_{h,j} = l \mid w_{h,l}) = w_{h,l}$ for $l = 1, \dots, P$ with the weights $w_{h,l}$ obtained through a stick-breaking construction (Sethuraman, 1994), i.e. $w_{h,j} = v_{h,j} \prod_{l=1}^{j-1}(1 - v_{h,l})$, $v_{h,j} \sim \mathrm{Beta}(\beta_1, \beta_2)$, $1 \leq j \leq P$. Hence, the probability of selecting the target spike is increasing with the lags j, since $P(z_{h,j} \leq j) = \sum_{l=1}^{j} w_{h,l}$. Correspondingly, the probability of choosing the Inverse Gamma slab component is $P(z_{h,j} > j) = \prod_{l=1}^{j}(1 - v_{h,l})$, i.e. decreasing with j. Higher sparsity levels for the modes $\alpha_{1,h}, \alpha_{2,h}$ and $\alpha_{3,h}$ are obtained by setting smaller values of a_τ and b_λ relative to b_τ and a_λ, respectively. We discuss these choices in Section 3.3.4.

3.3.3 Rank of the PARAFAC decomposition

A crucial point in the representation (3.3) is the choice of the rank, H. One widely adopted option is to regard this choice as a model selection problem and naturally resort to information criteria such as AIC or BIC (Zhou et al., 2013; Wang et al., 2021, 2016; Davis et al., 2016). As an alternative, we rely on results from recent Bayesian literature on overfitted mixture models (Malsiner-Walli et al., 2016; Rousseau and Mengersen, 2011) and set the parameters of the sparsity-inducing hierarchical prior in Section 3.3.2 so to automatically shrink unnecessary components to zero. Under quite general conditions, the posterior distribution concentrates on a sparse representation of the true density (Rousseau and Mengersen, 2011). More specifically, a small concentration parameter α of the symmetric Dirichlet distribution $(\phi_1, \ldots, \phi_H)$ will assign more probability mass to the edges of the simplex, meaning that more components become redundant. As a result, only a small number of components – within the H available – will be effectively different from zero. In addition, the cumulative shrinkage prior for the VAR lag order P can also be employed to encourage shrinkage, by choosing appropriate β_1, β_2, a_w and b_w. For instance, a large value of β_1 and a small β_2 encourage a more parsimonious VAR model by putting little probability on higher orders. Therefore, at the expense of a slightly higher computational demand, it is possible to fix relatively high values for H (and P), and then let the regularization implied by the shrinkage priors determine the number of effective components (lags), without the need for ranking different models in practice.

Figure 3.3 summarizes the proposed hierarchical model on the N-dimensional time series $\boldsymbol{y}_t$ in a directed graph representation.

3.3.4 Posterior Computation

In order to conduct inference on the dynamic coefficient matrices $\left[A_{1,t}, A_{2,t}, \ldots, A_{P,t}\right]$ and the latent indicators $\gamma_{h,t}$ we need to revert to the use of Markov Chain Monte Carlo methods. The prior specification allows to use a blocked Gibbs sampler to draw samples from the posterior distribution. When sampling from the posterior distribution for θ_h and κ_h using a Metropolis-Hastings algorithm, the normalizing constant depends on the sampled parameters, giving rise to a well-known issue of sampling from a doubly-intractable distribution. Thus, we follow the auxiliary variable approach proposed by Møller et al. (2006) to obtain the posterior samples from the Ising model. More specifically, the ap-

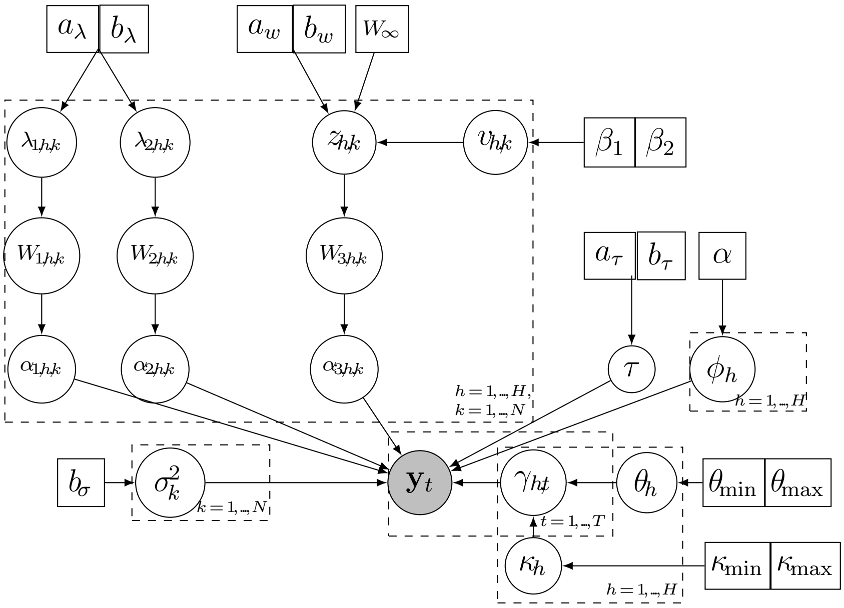


Figure 3.3: A schematic directed graph representation of the hierarchical model. In particular, the graph summarizes the Ising prior on the latent selection indicators $\gamma_{h,t}$ and the regularization priors on the tensor margins $\alpha_{1,h}$, $\alpha_{2,h}$, $\alpha_{3,h}$ of the rank-1 PARAFAC decomposition of the base components $A^*_{h,j}$, $h = 1, \ldots, H$, $j = 1, \ldots, P$ through

proach introduces an auxiliary variable such that – by adding its full conditional to the Metropolis-Hasting ratio – the normalizing constant is canceled out. Posterior samples of the latent auxiliary variable are then obtained using an exact sampling algorithm with coupled Markov chains (Propp and Wilson, 1996). The use of the auxiliary variable approach is also key for allowing the update of the $\gamma_{h,t}$'s for all $t = P + 1, \ldots, T$ and each $h = 1, \ldots, H$. The details of the MCMC and the full conditional distributions are reported in the appendix.

After obtaining posterior samples via MCMC, we can obtain inferences on the dynamic coefficient matrices $\left[A_{1,t}, A_{2,t}, \ldots, A_{P,t}\right]$. We compute the posterior means at each time point by averaging the values sampled across the MCMC iterations. We summarize inference on the $\gamma_{h,t}$'s by computing the posterior mode, i.e. we set $\tilde{\gamma}_{h,t} = 1$ whenever the posterior probability of activation, $P(\gamma_{h,t} = 1 \mid y_1, \ldots, y_T)$, is over 0.5.

3.4 Simulation Studies

In the first simulation study, we aim to evaluate whether our approach recovers the true dynamic coefficients under a time-varying VAR model. We generate 100 different datasets where the VAR coefficients are randomly generated from Gaussian distributions, according to the low rank tensor structure (3.2) together with random binary variables indicating the dynamic process. More specifically, we simulated 100 samples $(y_1, \ldots, y_T)$ from an $N = 10$-dimensional VAR model of order $P = 3$ and underlying tensor decomposition rank $H = 3$, with $T = 100$. In each sample, the noise terms of the 10 time series are assigned zero-centered Gaussian distributions, each with a different standard deviation, specifically equal to $1/5, 2/5, \ldots, 10/5$. In each data set, $\alpha_{1,h}, \alpha_{2,h}$ and $\alpha_{3,h}, h = 1, 2, 3$ are sampled from a spike-and-slab prior where the slab component is a standard normal distribution and the probability of a non-zero entry in $\alpha_{1,h}, \alpha_{2,h}$ and $\alpha_{3,h}$ is 0.5 (Mitchell and Beauchamp, 1988). We further ensure that the resulting TV-VAR time series are stationary. To sample the dynamic indicators, we generate $\gamma_{h,t}$, $h = 1, 2, 3$, $t = P+1, \ldots, 100$ from an NDARMA model whose parameters p_1 and p_2 follow a uniform distribution on $(0, 1)$.

Table 3.1: Simulation study 1. Bayesian point estimates (posterior means) of the identified tensor components and the dynamic coefficients from the proposed BTVT-VAR model. The latter are compared with the frequentist estimates of a time-varying VAR model implemented in the tvReg R-package. The evaluation of the tensor components is based on the square-root of the average Frobenius norm of the difference between the posterior mean and the true matrices across the 100 data sets, divided by number of entries. Columns 2 and 3 show the average Euclidean distances for each truly non-zero and truly zero entry in the matrices. Standard deviations are indicated in brackets. See Section 5 for details.

		All entries	True non-zero entries	True zero entries
BTVT-VAR	PARAFAC Components $A^*_{j,h}$	0.0361 (0.0335)	0.1190 (0.1557)	0.0130 (0.0165)
	VAR Coefficients Matrices $A_{j,t}$	0.0657 (0.0413)	0.1526 (0.1199)	0.0270 (0.0202)
TvReg	VAR Coefficients Matrices $A_{j,t}$	0.3967 (0.3853)	0.4550 (0.5265)	0.3815 (0.3650)

For model fitting, we assume mis-specified values of $H = 4$ and $P = 4$ in order

Table 3.2: Simulation study 1. Performance evaluation of the posterior estimation of the components' indicators $(\gamma_{h,P+1},\dots,\gamma_{h,T})$, based on average accuracy, sensitivity, specificity and precision across the 100 generated data sets. Standard deviations are indicated in brackets. See Section 5 for details.

	Accuracy	Sensitivity	Specificity	Precision
$(\gamma_{h,P+1},\dots,\gamma_{h,T})$	0.9018 (0.1333)	0.9634 (0.0881)	0.6459 (0.3741)	0.9049 (0.1272)

to assess our method's ability to automatically determine the true rank and lag-order of the VAR model. To encourage sparsity, we set the hyper-parameter values as $a_\lambda = 3, b_\lambda = \sqrt[6]{a_\lambda} = \sqrt[6]{3}, b_\tau = 4^{-2/3}, \beta_1 = 1, \beta_2 = 5, a_w = b_w = 2, W_\infty = 0.01, a_\sigma = b_\sigma = 1$, $\theta_{h,\min} = -4, \theta_{h,\max} = 4, \kappa_{h,\max} = 4$. For the griddy-Gibbs step of the posterior sampling, α is assumed to be uniformly distributed across 10 values evenly spaced in the interval $[H^{-3}, H^{-0.1}]$ (Guhaniyogi et al., 2017). Finally, a total of 5,000 MCMC iterations are run, one third of the output is discarded and the remaining samples are thinned by a factor of 3 to reduce storage and possible auto-correlation of the chains.

We first summarize the results of the simulation study by investigating the ability of our model to recover the dynamic coefficient matrices $\left[A_{1,t}, A_{2,t}, \dots, A_{P,t}\right]$ in (3.2). To assess the performance of the proposed method, we employ the root mean square Frobenius distance between the MCMC estimates of the posterior means $E\left(A_{j,t} \mid y_1,\dots,y_T\right)$ at each t, say $\{\tilde{A}_{j,t}\}_{j=1,\dots,P,t=P+1,\dots,T}$, and the true matrices as

$$\operatorname{err}(\{\tilde{A}_{j,t}\}) = \sqrt{\frac{\sum_{t=P+1}^{T}\sum_{j=1}^{P}\|\tilde{A}_{j,t} - A_{j,t}\|_F}{(T-P)\times N^2P}}. \tag{3.7}$$

We also compare the obtained Bayesian point estimates with those provided by a frequentist time-varying vector auto-regressive model, as implemented in the R package TvReg (Casas and Fernandez-Casal, 2021, 2019). The results are shown in Table 3.1. The Bayesian time-varying tensor VAR (BTVT-VAR) model appears to provide an improved estimation of the true dynamic structure of the data with respect to the non-sparse frequentist VAR. Table 1 also shows the point estimates of the the base matrices $[A^*_{1,h}, A^*_{2,h}, \dots, A^*_{P,h}]$. The error of the MCMC-based estimates of the posterior means, say $\{\tilde{A}^*_{j,h}\}_{j=1,\dots,P}$, is similarly defined as

$$\text{err}(\{\tilde{A}^*_{j,h}\}) = \sqrt{\frac{\sum_{j=1}^{H} \|\tilde{A}^*_{j,h} - A^*_{j,h}\|_F}{N^2 P}}. \tag{3.8}$$

In order to compute (3.8), since the estimation of the components of the mixture (3.2) may be affected by label switching (Stephens, 2000), it is necessary first to match the posterior means of each component to the true components. In addition, if the assumed value of H is larger than the true value, one or more of the posterior mean estimates $\{\tilde{A}^*_{j,h}\}_{j=1,\dots,P}$ could likely be redundant and include elements all very close to zero. In order to identify these essentially "empty" components, we consider the maximum norm $\max_{j=1,\dots,P} \|\tilde{A}^*_{j,h}\|_\infty$ and if such norm is lower than a pre-specified threshold, we set them to $\mathbf{0}$ and exclude them from further analysis. More specifically, in the following, a component is assumed as empty if its posterior mean-based maximum norm is smaller than 0.01. Then, in order to match the remaining posterior mean estimates $\{\tilde{A}^*_{j,h}\}_{j=1,\dots,P}$ with the true components, we rank them based on the minimum Frobenius distances.

For the inference on the latent binary indicators $(\gamma_{h.P+1}, \dots, \gamma_{h.T})$, $h = 1, \dots, H$, we threshold the estimated posterior probability of activation at each time point for each data set to identify the activated $\gamma_{h,t}$'s from the MCMC samples, as described in Section 3.3.4. Table 3.2 shows the average accuracy, sensitivity, specificity and precision across all 100 data sets. The results show that the model is able to reconstruct the components and their dynamic activation reasonably well. Further inspection of the results across all simulated data sets suggests – as it may be expected – that the ability to identify each underlying base mixture component is associated to the sparsity of the true activation vectors $(\gamma_{h,P+1}, \dots, \gamma_{h,T})$, $h = 1, \dots, H$ (analyses not shown). Barely activated components are more difficult to identify as it is more challenging to disentangle their contribution,

Table 3.3: Simulation study 2. Bayesian point estimates (posterior means) of the identified tensor components and dynamic coefficients from the proposed BTVT-VAR model. The latter are compared with the frequentist estimates of a time-varying VAR. model implemented in the tvReg R-package. The evaluation of the tensor components is based on the square-root of the average Frobenius norm of the difference between the posterior mean and the true matrices across three true components. Columns 2 and 3 show the average Euclidean distances for each truly non-zero and truly zero entry in the matrices. Standard deviations across the three true components are indicated in brackets. See Section 5 for details.

		All entries	True non-zero entries	True zero entries
BTVT-VAR	PARAFAC Components $A^*_{j,h}$	0.0127 (0.0009)	0.0588 (0.0520)	0.0083 (0.0023)
	VAR Coefficients Matrices $A_{j,t}$	0.1083	0.2538	0.0694
TvReg	VAR Coefficients Matrices $A_{j,t}$	0.3798	0.3056	0.3885

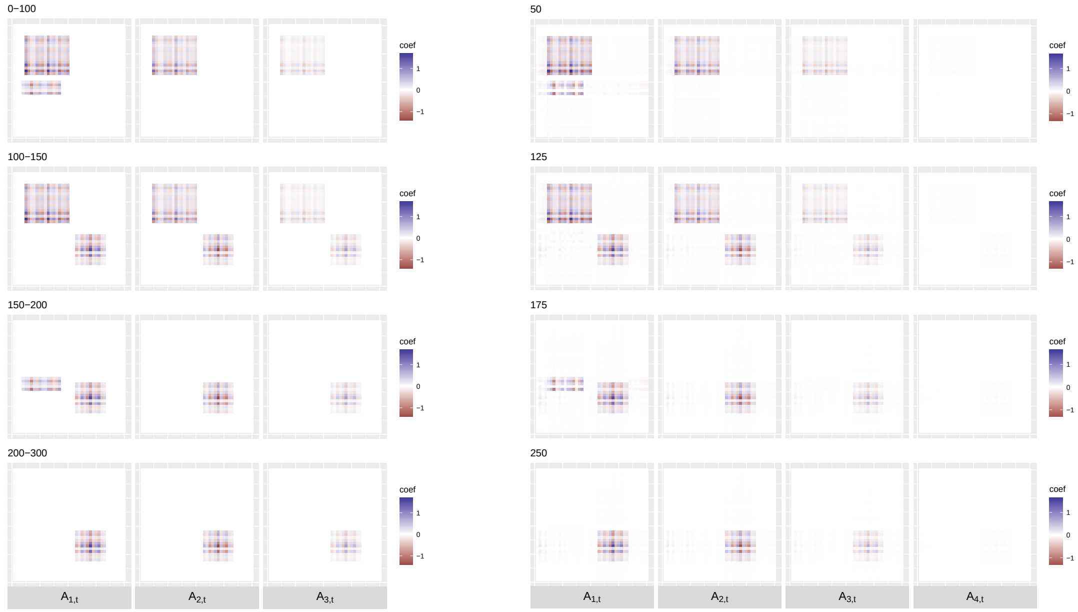
Table 3.4: Simulation study 2. Performance evaluation of the posterior estimation of the components' indicators $(\gamma_{h,P+1}, \dots, \gamma_{h,T})$. The evaluation is based on average accuracy, sensitivity, specificity and precision over the 3 true components.

	Accuracy	Sensitivity	Specificity	Precision
$(\gamma_{h,P+1}, \dots, \gamma_{h,T})$	0.9887 (0.0195)	0.9886 (0.0198)	0.9889 (0.0192)	0.9886 (0.0198)

especially if the magnitude of their coefficients (their "signal") is low. Also, the relative dynamics of the components, i.e. how they differentially activate in time, has an impact on identifiability. Finally, we evaluate the performance of the proposed shrinkage priors to determine the rank of the tensor decomposition as well as the lags in the TVT-VAR model. Since we use $P = 4$ in model fitting although the true value is $P = 3$, we observe that the estimated $N * N$ matrix $A_{P,t}$ has on average a Frobenius norm of 0.0005 with 0.0015 standard deviations across all t's, suggesting that the proposed increasing shrinkage prior allows to identify the number of lags well.

In addition to simulating TVT-VAR coefficients from a spike-and-slab prior with some artificial normal slab distribution, we also consider a scenario where the time series are generated from TVT-VAR models whose coefficients are estimated from real fMRI data. The fMRI data that we use are the first two runs of participant 2 when reading Chapter 9 of Harry Potter and the Sorcerer's Stone. See Section 3.5 for more details of the data set. We first obtain the ordinary least square coefficient estimates of the VAR model with $N = 27, P = 3$ by fitting the two runs separately. For each estimated coefficient matrix, we stack it and apply a rank-2 PARAFAC decomposition. The resulting four different tensor margins $\alpha_{1,h}, \alpha_{2,h}, \alpha_{3,h}$ are then used to construct the time-varying coefficients. In total, 100 repetitions are done and repetition-specific tensor margins are samples from normal distributions with mean $\alpha_{1,h}, \alpha_{2,h}, \alpha_{3,h}$ obtained from real fMRI data as described above. Repetition-specific tensor margins are further thresholded by 0.1 to induce a desirable level of sparsity. Each repetition simulates a multivariate time course of length $T = 400$ from a TV-VAR model with coefficients changing at time points 80, 160, 240, 320. Inside each time interval, the coefficients are assumed to be the sum of a time-invariant subset of the four base matrices calculated from the tensor margins. Entries of the error term $\boldsymbol{\epsilon}_t$ have identical standard deviation determined by the signal-to-noise ratio (SNR). Many definitions of SNR exist in the literature (Welvaert and Rosseel, 2013). For instance, Zhang et al. (2014) define the SNR through the variance of the regression parameters and the innovation variance of the error term, whereas Haslbeck et al. (2021) use the ratio between the maximum parameter size of time-varying parameters and the noise variance. We follow the definition of Haslbeck et al. (2021) and divide the maximum size of the estimated VAR coefficients by the estimated error term standard deviation to obtain the SNR value of 1.8518 and 2.7972 for the two runs, therefore we set the SNR equal to 2.5 to calculate the variance of the error term in simulation. To test the robustness of the proposed Bayesian model, we include two extra situations where the SNR is equal to 0.5 and 10 respectively. This is still in progress and the results are not reported here for the moment.

In the third simulation study, we move to higher dimensional models and generate a time series from an $N = 40$ TV-VAR model of order 3 with dynamic coefficients shown in the left panel of Figure 3.4. These coefficients are combinations of $H = 3$ components. A total of $T = 300$ observations are simulated, among which the coefficients matrix of the first 200 observations admits a rank-2 tensor decomposition with changing mixing components whereas the last 100 consist of only one component. The covariance matrix Σ of the error term ϵ_t has diagonal elements $\Sigma[i,i] = i/5$ for $i = 1, \ldots, 25$ and $\Sigma[i,i] = (51 - i)/5$ for $i = 26, \ldots, 40$. In the posterior inference, we choose $H = 4$ and $P = 4$. Even though we include one more component and lag, we expect the extra component as well as coefficients at lag larger than 3 to be almost zero due to the effect of the shrinkage priors. The remaining experiment settings are the same as in the first simulation. We select four estimated coefficients matrices at time point 50, 125, 175 and 250 as in the right panel of Figure 3.4. Our method identifies three "non-empty" components and they accurately capture the patterns of the true ones. Furthermore, the dynamics of the coefficient matrices are all accurately identified by the model. Table 3.3 shows an evaluation of the posterior inference on the coefficient matrices in our model versus a frequentist time-varying regression. Once again, our model compares quite favorably. To further verify our method's ability to detect changing patterns along all the 296 time points, we report the accuracy, sensitivity, specificity and precision in the estimation the latent activation indicators $\gamma_{h,t}$'s in Table 3.4. In Figure 3.5 we report the estimated trajectories of the $\gamma_{h,t}$ as a function of time, for each component $h = 1, \ldots, H$. The shaded red areas indicate the true component activation, whereas the solid line indicates the posterior modes. One of the trajectories, being constantly zero, identifies an empty component: indeed, we employed $H = 4$ for model fitting instead of the true number of components. The other three estimated trajectories follow the true activations quite closely, reaching false positive rates of 0%, 0.67% and 0% as well as false negative rate of 0%, 3.42% and 0%, respectively. Once again, the results illustrate the role of the shrinkage prior specifications, since fixing higher values of H and P does not appear to hamper the estimation of the VAR matrices. In particular, we do not need to rely on model selection techniques in order to determine the appropriate values of H and P. Therefore, in many cases it may be desirable to learn the actual dimensions of the model



(a) Truth

(b) $E(A_{p,t}|y)$

Figure 3.4: Comparison between the true TVT-VAR coefficients $A_{1,t}$, $A_{2,t}$, $A_{3,t}$ on the left panel and the MCMC posterior means from the BTV-TVAR model on the right panel in the second simulation study of Section 5. For each truly invariant time window, the estimated matrices at time point 50, 150 and 250 are displayed.

from the data, by fixing relatively large values of H and P. As a final remark, we note that we also considered a simulation scenario with $N = 100$ multivariate time series. However, the frequentist TvReg approach did show numerical problems with such a large number of times series, after taking 9.44 hours to complete. On the contrary, our method was still able to obtain good inferences for this large dimensions, after taking 3.6 hours to complete 5,000 iterations on a Intel Core i5-6300U CPU at 2.40GHz, with 8GB RAM. As a comparison, each run of the N =40 -dimensional case took only approximately 56 minutes to complete for our method.

3.5 Real Data Application

We apply our TVT-VAR model to the following task-based functional magnetic resonance (fMRI) data set. This data includes 8 participants (ages 18–40), who were asked to read

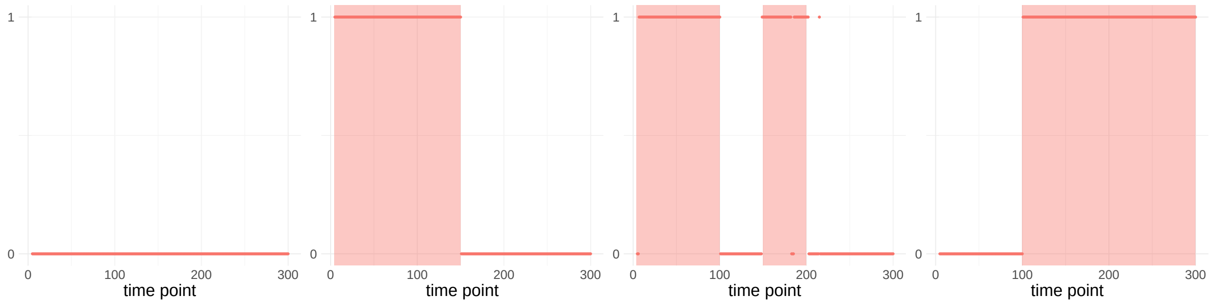


Figure 3.5: Estimated trajectories of the latent indicators of activation $\gamma_{h,t}$ for four components $h = 1, \ldots, 4$. The red areas indicate the true activations. The solid line indicates the estimated values of $\gamma_{h,t}$ based on posterior probabilities of activation greater than 0.5.

Chapter 9 of *Harry Potter and the Sorcerer's Stone* (Rowling, 2012). All subjects had previously read the book or seen the movie. The words of the story were presented in rapid succession, where each word was presented one by one at the center of the screen for 0.5 seconds in black font on a gray background. A Siemens Verio 3.0T scanner was used to acquire the scans, utilizing a T2* sensitive echo planar imaging pulse sequence with repetition time (TR) of 2s, time echo (TE) of 29 ms, flip angle (FA) of 79°, 36 number of slices and $3 \times 3 \times 3$mm^3 voxels. Data was pre-processed in the following manner. For each subject, functional data underwent realignment, slice timing correction, and co-registration with the subject's anatomical scan, which was segmented into grey and white matter and cerebro-spinal fluid. The subject's scans were then normalized to the MNI space and smoothed with a $6 \times 6 \times 6$mm Gaussian kernel smoother. Data was then detrended by running a high-pass filter with a cut-off frequency of 0.005Hz after being masked by the segmented anatomical mask. The final time series for the task-based data contained 4 runs for each subject. Runs 1, 2, 3, and 4 contained 324, 337, 264, and 365 time points, respectively. For more details, see Ondrus et al. (2021) and Xiong and Cribben (2021).

Twenty-seven ROIs defined using the Automated Anatomical Atlas (AAL) brain atlas (Tzourio-Mazoyer et al., 2002) were extracted from the data set, shown in Table 3.5. These regions contain a variety of voxels that have been previously recognized as important to

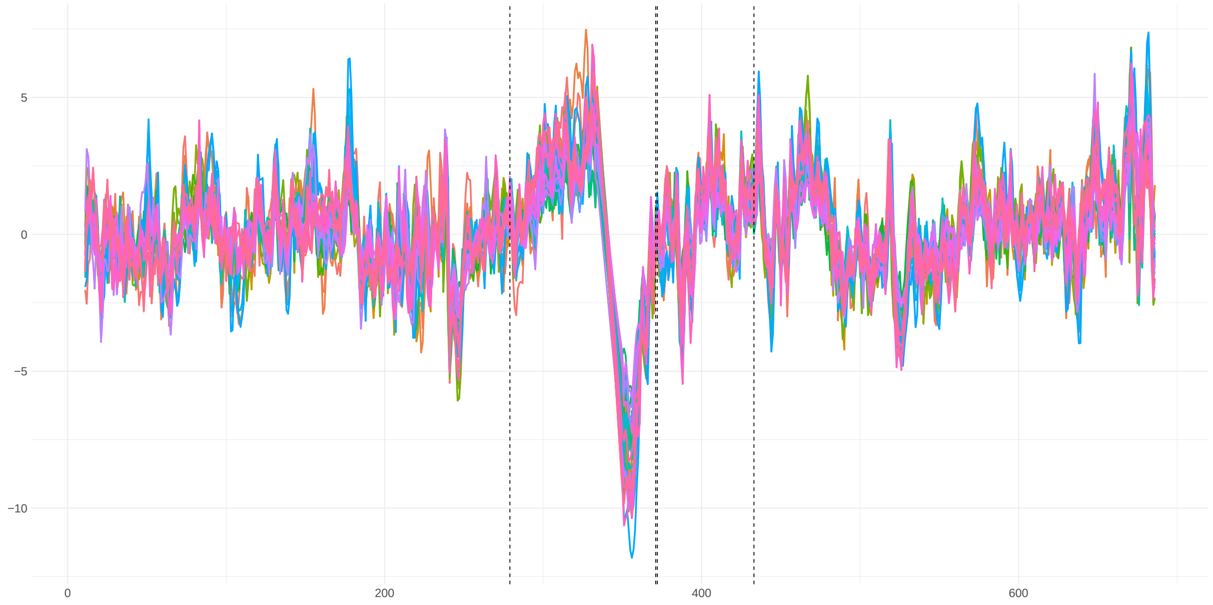


Figure 3.6: fMRI recording for a representative subject in the application to the fMRI reading experiment described in Section 3.5. The dashed vertical lines on the left and the right represent the time windows considered before and after Harry Potter's first flying experience on a broom. See Section 3.5 for details.

distinguish between the literary content of two novel text passages based on neural activity while these passages are being read. In this work, we focus on exploring the dynamic effective connectivity of the $N = 27$ ROIs as the subjects read using the proposed BTVT-VAR model. More specifically, for model fitting, we choose $P = 4$. Most applications of VAR models to fMRI data consider $AR(1)$ processes as a good representation of short-range temporal dependences over a small number of regions of interest (typically, less than ten). We further choose $H = 10$ as the number of mixture components in (3.2). This value was chosen as it allows to recover more than 50% of the estimated variability in the sample, as assessed by the Frobenious norm of the coefficient matrices estimated in a frequentist VAR LASSO. We are interested in the varying textual features about story characters (e.g., emotion, motion and dialog) that the dynamic effective connectivity encode.

As reading is a complex task, we focus on how our BTVT-VAR model responds to significant plot changes. For example, around time point $t = 368$, Harry has his first

Table 3.5: Information on the 27 ROIs extracted from the Harry Potter task-based fMRI data set. Each ROI has a left and right hemisphere component apart from the Supramarginal gyrus.

ROI id	**Regions**	**Label**
1	Angular gyrus	AG
2	Fusiform gyrus	F
3	Inferior temporal gyrus	IT
4	Inferior frontal gyrus, opercular part	IFG 1
5	Inferior frontal gyrus, orbital part	IFG 2
6	Inferior frontal gyrus, triangular part	IFG 3
7	Middle temporal gyrus	MT
8	Occipital lobe	O
9	Precental gyrus	PCG
10	Precuneus	PC
11	Supplementary motor area	SM
12	Superior temporal gyrus	ST
13	Temporal pole	TP
14	Supramarginal gyrus	SG.R

flying experience on a broom. Hence, we estimate the BTVT-VAR on time points before and after this event. More specifically, we combine runs 1 and 2 of each individual, and summarize the time-varying coefficient matrices by averaging the posterior means within specific time windows before and after the reading of the flying experiences respectively (the specific time may vary slightly for different individuals), i.e. $\hat{A}_j = \sum_{t \in W} \hat{A}_{j,t}/|W|$, for each time window W, with $|W|$ indicating the length of the window, $\hat{A}_{j,t}$ the MCMC estimate of the posterior mean at time t and $j = 1, \ldots, P$. Figure 3.6 shows the blood-oxygen-level-dependent (BOLD) fMRI data in a representative subject, with the corresponding windows considered before and after the flying experience. The number of parameters considered in the fitted BTVT-VAR is $H(T - P) + H(2N + P) = 7,150$, versus $(T - P)N^2P = 1,915,812$ of a standard time-varying VAR model.

Figure 3.7 shows the estimated coefficients in the BTVT-VAR model before (top panel) and after (middle panel) the flying experience as well as the difference (bottom panel) between the mean coefficients before and after for subjects 1, 3, and 5. We only plot the lag-1 mean coefficients, $(\hat{A}_1)$ as the other lag mean coefficient matrices have very little activity. Overall, the signs of the coefficients by-and-large coincide across the subjects. There are also clear common patterns across the subjects. For the before coefficient matrices (Figure 3.7, top row), all subjects have large positive and negative coefficients between the left and right occipital lobe (O.L and O.R) and all other ROIs. This is unsurprising given the its primary role is to provide the sense of vision and extracts information about the visual world, which is then passed on to other brain areas that mediate awareness. Another function includes movement. Furthermore, all subjects have moderate positive coefficients between the fusiform gyrus (F.L and F.R) and all other ROIs. The fusiform gyrus plays important roles in object and face recognition, and recognition of facial expressions is located in the fusiform face area (FFA), which is activated in imaging studies when parts of faces or pictures of facial expressions are presented (Kleinhans et al., 2008). There is also evidence that this brain region plays a role in early visual processing of written words (Zevin, 2009). There is also some heterogeneity across the subjects. For example, subjects 1 and 3 have large positive coefficients between the middle temporal gyrus (MT.L but mostly for MT.R) and a large potion of the other ROIs. The middle temporal gyrus is sensitive to visual motion (flying), and while traditional language processing areas include the inferior frontal gyrus (Broca's area), superior temporal and middle temporal gyri, supramarginal gyrus and angular gyrus (Wernicke's area), there is evidence that structures in the medial temporal lobe have a role in language processing (TRACY and BOSWELL, 2008). Subject 5 also shows signs of heterogeneity. For example, unlike subjects 1 and 3, its connectivity patterns reveal large positive coefficients between the right temporal pole (TP.R) and a large potion of the other ROIs. The temporal pole has been associated with several high-level cognitive processes: visual processing for complex objects and face recognition, naming and word-object labelling, semantic processing in all modalities, and socio-emotional processing (Herlin et al., 2021).

The structures in the after coefficient matrices (Figure 3.7, middle row) are overall quite similar to the before coefficient matrices (Figure 3.7, top row) indicating a smooth transition over this period of the book. However, there are some differences which are depicted in Figure 3.7 (bottom row). Subjects 1 and 5 have strong negative coefficients between the right angular gyrus (AG.R) and most of the other ROIs. The angular gyrus is known to participate to complex cognitive functions, such as calculation (Duffau, 2012). The angular gyrus, especially in the right hemisphere, is essential for visuospatial awareness. These regions may generate the fictive dream space necessary for the organized hallucinatory experience of dreaming (Pace-Schott and Picchioni, 2017). The sequence of events occurring in the book at this time require both calculation and imagination for picturing how flying on a broom would materialize. Additionally, subjects 3 and 5 have moderate positive coefficients between the right precuneus gyrus (PCG.R) and most of the other ROIs. The precuneus is involved in a variety of complex functions, which include recollection and memory, integration of information relating to perception of the environment, cue reactivity, mental imagery strategies, episodic memory retrieval (Borsook et al., 2015). Moreover, the strong relationship between the medial temporal lobe and precuneus, and the referred circuitry connecting these areas is referred to as the default mode network (Greicius et al., 2003). There is also heterogeneity in the differences across subjects. For example, subject 1 has large positive coefficients between the right supramarginal gyrus (SG.R) and most of the other ROIs. Similar to the angular gyrus, the right supramarginal gyrus is essential for visuospatial awareness and it may generate the fictive dream space necessary for the organized hallucinatory experience of dreaming (Pace-Schott and Picchioni, 2017).

Another plot twist occurs close to time point $t = 1176$ in run 4. Here, Harry, Ron and Hermione (the main three characters in the book), arrive in a forbidden corridor, turn around and come face-to-face with a monstrous three-headed dog. This event is the most thrilling in Chapter 9 in *Harry Potter and the Sorcerer's Stone.* Hence, we estimate the BTVT-VAR on time points before and after this event. Figure 3.8 shows the estimated coefficients in the BTVT-VAR model before (top panel) and after (middle panel) coming face-to-face with the dog as well as the difference (bottom panel) between the mean coefficients before and after for subjects 3, 7, and 8. We only plot the lag-1 mean coefficients ($\hat{A}_1$) as the other lag mean coefficient matrices have very little activity. Overall, in this example, there is a great deal of heterogeneity across the subjects. For the before coefficient matrices (Figure 3.8, top row), subject 3 has large coefficients between the occipital lobe (O.L and O.R) and all other ROIs, subject 7 has a moderately strong network between the inferior frontal gyrus, orbital part (IFG.2), inferior frontal gyrus, triangular part (IFG.3), inferior temporal gyrus (IT), precuneus (PC) and supplementary motor area (SM), and subject 8 has large coefficients between the left and the right precuneus gyrus (PCG.L and PCG.R) and most of the other ROIs.

As in the first example, the structures in the after coefficient matrices (Figure 3.8, middle row) are overall quite similar to the before coefficient matrices (Figure 3.8, top

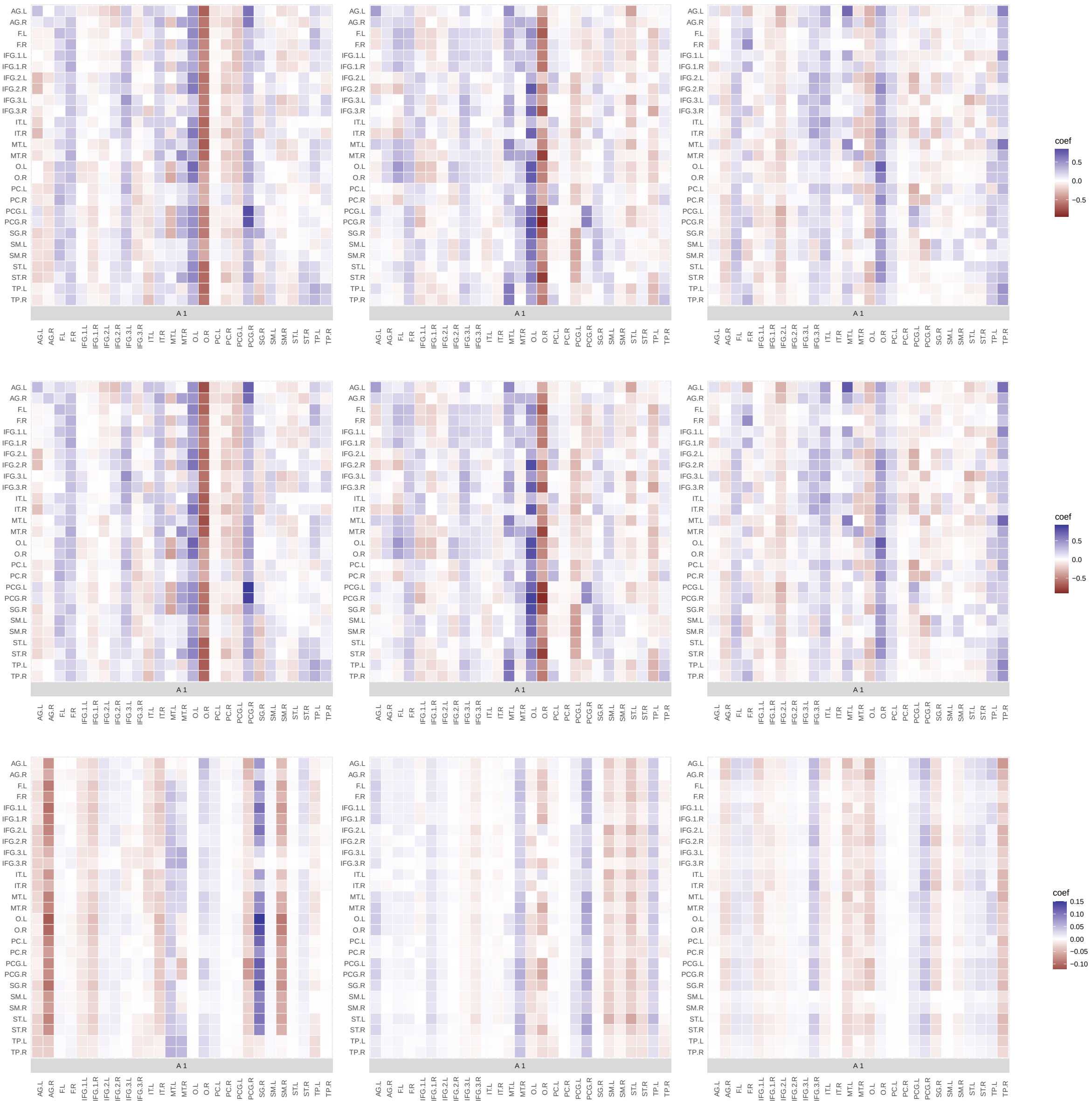


Figure 3.7: The estimated coefficients in the BTVT-VAR model before (top panel) and after (middle panel) Harry's first flying experience (around time point $t = 368$) as well as the difference (bottom panel) between the mean coefficients before and after for subjects 1 (first column), 3 (second column), and 5 (third column) in the Harry Potter fMRI data set. We only plot the lag-1 mean coefficients ($\hat{A}_1$). The 27 ROI names can be found in Table 3.5.

row) indicating a smooth transition over this period of the book. However, there are some differences which are depicted in Figure 3.8 (bottom row). Subject 8 has the most stark differences between the before and after. In particular, there are large negative and positive coefficients between the left and right precuneus gyrus (PCG.R) and all other ROIs, respctively. As mentioned above, the precuneus can be divided into regions involved in sensorimotor processing, cognition, and visual processing. Subject 8 also has large negative coefficients between the left and right temporal pole (TP.L and TP.R) and a large potion of the other ROIs. The temporal pole has been associated with several high-level cognitive processes: visual processing for complex objects and face recognition (Herlin et al., 2021). The visualization of the meeting with the three-headed dog would require a significant amount of visual processing for complex objects. Futhermore, subject 8 also has large positive coefficients between the right supramarginal gyrus and several of the other ROIs. The right supramarginal gyrus is essential for visuospatial awareness and it may generate the fictive dream space necessary for the organized hallucinatory experience of dreaming (Pace-Schott and Picchioni, 2017). The sequence of events occurring in the book at this time are dreamlike with the description including the word “nightmare” and the characters moving from a room to a corridor without their own movement.

The difference in the coefficients for subjects 3 and 7 are not as strong. Subject 3 has differences in the coefficients between the left occipital lobe (sense of vision and extracts information about the visual world), the inferior frontal gyrus andopercular part (language processing) and many of the other ROIs. Subject 7 has differences in the coefficients between the left angular gyrus (complex cognitive functions) and some of the other ROIs and between the left fusiform gyrus and almost all the ROIs. The fusifor gyrus is involved in the processing the printed forms of words (Zevin, 2009).

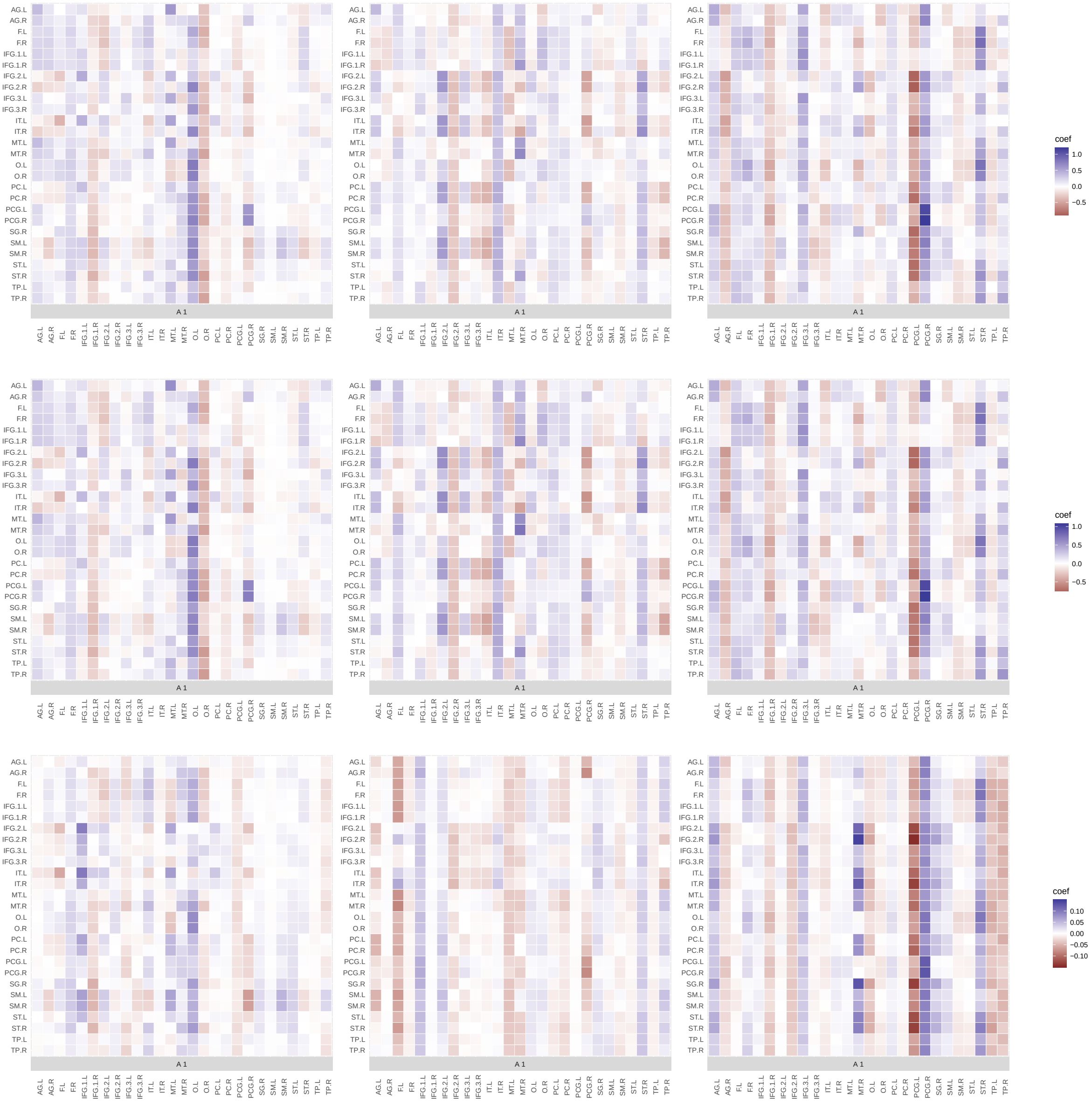


Figure 3.8: The estimated coefficients in the BTVT-VAR model before (top panel) and after (middle panel) Harry, Ron and Hermione (the main three characters in the book), arrive in a forbidden corridor, turn around and come face-to-face with a monstrous three-headed dog (around time point $t = 1176$) as well as the difference (bottom panel) between the mean coefficients before and after for subjects 3 (first column), 7 (second column), and 8 (third column) in the Harry Potter fMRI data set. We only plot the lag-1 mean coefficients ($\hat{A}_1$). The 27 ROI names can be found in Table 3.5.

3.6 Discussion and future work

We have proposed a scalable Bayesian time-varying tensor VAR model for the study of effective connectivity in fMRI experiments, and we have shown that it results in good performances and interpretable results in both a simulation and an application to a dataset from a complex text reading experiment.We focus on applications to fMRI data, where an AR(1) dependence is often assumed as sufficient (however, see Monti, 2011; Corbin et al., 2018, for different takes). Our data analysis appears to confirm this general suggestion, as higher lags of the coefficient matrices did not show patterns relevantly different from zero in our fMRI experiment. However, our method is applicable to other types of time-varying neuroimaging data, including EEG data, where higher orders of auto-regression are more natural, and more generally to any type of data where a vector autoregressive model is appropriate.

An important feature of the proposed time-varying tensor VAR model is that it implicitly relies on a state-space representation, with the state space containing 2^H elements. A state is obtained as the composition of a subset of the H components shared over the entire time span. This representation could be leveraged to obtain scalable inference and describe shared patterns of brain connectivity in multi-subject analyses. More specifically, in addition to allowing a more parsimonious representation of the coefficient matrices than required by traditional non-tensor approaches, our formulation could be employed to identify temporally persistent connectivity patterns in some brain areas, by tracking the components that remain active over multiple time intervals and multiple subjects. However, this type of inference would require allowing the identifiability of the same tensor components across subjects. One way to achieve this result is through the use of clustering-inducing Bayesian nonparametric priors, that will allow also borrowing of information across all subjects in estimating the components. Due to the increased computational burden this solution will require, we leave its exploration of these avenues to future work.

Appendix

Proof of Proposition 1

Here we prove the equivalence between the Ising prior and the NDARMA(1) model, which relies on the exponential formulation of the multivariate Bernoulli distribution in Dai et al. (2013).

Proof. We use mathematical induction to prove the theorem. First, the probability mass function of the NDARMA(1) model with parameter (p_1, p_2) takes the general form of multivariate Bernoulli distribution as

$$\mathrm{pr}(\gamma_1, \dots, \gamma_T) \propto \exp\left(\sum_{n}^{T}\left(\sum_{1\le j_1<j_2<\cdots<j_n\le T} f_T^{j_1 j_2 \cdots j_n}\, B^{j_1 j_2 \cdots j_r}(\boldsymbol{\gamma})\right)\right).$$

When $T = 2$,

$$\exp(f_2^1) = \frac{p_{10}}{p_{00}} = \frac{p_2(1-p_1)}{p_1 + (1-p_2)(1-p_1)}, \quad \exp(f_2^2) = \frac{p_{01}}{p_{00}} = \frac{p_2(1-p_1)}{p_1 + (1-p_2)(1-p_1)},$$

$$\exp(f_2^{12}) = \frac{p_{00}p_{11}}{p_{10}p_{01}} = \frac{p_1 + p_2(1-p_2)(1-p_1)^2}{p_2(1-p_2)(1-p_1)^2}.$$

For any $T \ge 3$, it holds that when $j_n < T-1$,

$$\begin{aligned}\exp(f_T^{j_1\cdots j_n}) &= \prod_{\{\text{even } \# \text{ of 0's in } \gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T,j_1\cdots j_n} \Bigg/ \prod_{\{\text{odd } \# \text{ of 0's in } \gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T,j_1\cdots j_n}\\ &= \prod_{\{\text{even } \# \text{ of 0's in } \gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T-1,j_1\cdots j_n} \Bigg/ \prod_{\{\text{odd } \# \text{ of 0's in } \gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T-1,j_1\cdots j_n}\\ &= \exp(f_{T-1}^{j_1\cdots j_n}).\end{aligned}$$

From this equation, we can induce, in the case where $T = 3$, that

$$\exp(f_3^1) = \exp(f_2^1) = \frac{p_2(1-p_1)}{p_1 + (1-p_2)(1-p_1)}.$$

When $j_n = T - 1$,

$$\begin{aligned}
&\exp(f_T^{j_1\dots j_n})\\
=&\prod_{\{\text{even \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T,j_1\dots j_n}\Big/\prod_{\{\text{odd \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T,j_1\dots j_n}\\
=&\prod_{\substack{\{\gamma_{j_n}=0,\text{ odd \# of}\\ \text{0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}}} p^*_{T-1,j_1\dots j_n}(p_1+(1-p_1)(1-p_2)) \prod_{\substack{\{\gamma_{j_n}=1,\text{ even \# of}\\ \text{0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}}} p^*_{T-1,j_1\dots j_n}(1-p_1)(1-p_2)\Big/\\
&\prod_{\substack{\{\gamma_{j_n}=1,\text{ odd \# of}\\ \text{0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}}} p^*_{T-1,j_1\dots j_n}(1-p_1)(1-p_2) \prod_{\substack{\{\gamma_{j_n}=0,\text{ even \# of}\\ \text{0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}}} p^*_{T-1,j_1\dots j_n}(p_1+(1-p_1)(1-p_2))\\
=&\prod_{\{\text{even \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T-1,j_1\dots j_n}\Big/\prod_{\{\text{odd \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_n}\}} p^*_{T-1,j_1\dots j_n}\\
&\prod_{\{\text{even \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}} (1-p_1)(1-p_2)\Big/\prod_{\{\text{odd \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}} (1-p_1)(1-p_2)\\
&\prod_{\{\text{odd \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}} p_1+(1-p_1)(1-p_2)\Big/\prod_{\{\text{even \# of 0's in }\gamma_{j_1},\dots,\gamma_{j_{n-1}}\}} p_1+(1-p_1)(1-p_2)
\end{aligned}$$

It is equal to $e^{f_{T-1}^{j_1\dots j_n}}$ if $n > 1$ and $e^{f_{T-1}^{j_1\dots j_n}}(1-p_1)(1-p_2)/\left(p_1 + (1-p_1)(1-p_2)\right)$ if $n = 1$.

For instance,

$$\exp(f_3^{12}) = \exp(f_2^{12}) = \frac{p_1 + p_2(1-p_2)(1-p_1)^2}{p_2(1-p_2)(1-p_1)^2},$$

$$\exp(f_3^2) = \exp(f_2^2)\frac{(1-p_1)(1-p_2)}{p_1 + (1-p_1)(1-p_2)} = \frac{p_2(1-p_2)(1-p_1)^2}{(p_1 + (1-p_2)(1-p_1))^2}.$$

In the last case where $j_n = T$, depending on the value of j_{n-1}, we have

$$\begin{aligned}
&\exp(f_T^{j_1 \dots j_n}) \\
= &\prod_{\{\text{even \# of 0's in } \gamma_{j_1}, \dots, \gamma_{j_n}\}} p^*_{T,j_1 \dots j_n} \Bigg/ \prod_{\{\text{odd \# of 0's in } \gamma_{j_1}, \dots, \gamma_{j_n}\}} p^*_{T,j_1 \dots j_n} \\
= &\prod_{\substack{\{\gamma_{j_n}=0, \text{ odd \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T,j_1 \dots j_n} \prod_{\substack{\{\gamma_{j_n}=1, \text{ even \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T,j_1 \dots j_n} \Bigg/ \\
&\prod_{\substack{\{\gamma_{j_n}=1, \text{ odd \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T,j_1 \dots j_n} \prod_{\substack{\{\gamma_{j_n}=0, \text{ even \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T,j_1 \dots j_n} \\
= &\prod_{\substack{\{\gamma_{j_n}=0, \text{ odd \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T-1,j_1 \dots j_{n-1}}(p_1 + (1-p_1)(1-p_2)) \prod_{\substack{\{\gamma_{j_n}=1, \text{ even \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T-1,j_1 \dots j_{n-1}}(1-p_1)p_2 \Bigg/ \\
&\prod_{\substack{\{\gamma_{j_n}=1, \text{ odd \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T-1,j_1 \dots j_{n-1}}(1-p_1)p_2 \prod_{\substack{\{\gamma_{j_n}=0, \text{ even \# of 0's} \\ \text{in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}}} p^*_{T-1,j_1 \dots j_{n-1}}(p_1 + (1-p_1)(1-p_2)) \\
= &\prod_{\{\text{even \# of 0's in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}} (1-p_1)p_2 \Bigg/ \prod_{\{\text{odd \# of 0's in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}} (1-p_1)p_2 \\
&\prod_{\{\text{odd \# of 0's in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}} p_1 + (1-p_1)(1-p_2) \Bigg/ \prod_{\{\text{even \# of 0's in } \gamma_{j_1}, \dots, \gamma_{j_{n-1}}\}} p_1 + (1-p_1)(1-p_2)
\end{aligned}$$

when $j_{n-1} < T-1$. It takes value 1 when $n > 1$ and $p_2(1-p_1)/\left(p_1 + (1-p_2)(1-p_1)\right)$ when $n = 1$, which implies that

$$\exp(f_3^{13}) = 1, \quad \exp(f_3^3) = \frac{p_2(1-p_1)}{p_1 + (1-p_2)(1-p_1)}$$

in the $T = 3$ example.

Otherwise when $j_{n-1} = T - 1$,

$$\exp(f_T^{j_1 \ldots j_n})$$

$$= \prod_{\{\text{even \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_n}\}} p^*_{T,j_1\ldots j_n} \Big/ \prod_{\{\text{odd \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_n}\}} p^*_{T,j_1\ldots j_n}$$

$$= \prod_{\substack{\{\gamma_{j_{n-1}}=1,\\ \gamma_{j_n}=0, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n} \prod_{\substack{\{\gamma_{j_{n-1}}=0,\\ \gamma_{j_n}=1, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n} \prod_{\substack{\{\gamma_{j_{n-1}}=0,\\ \gamma_{j_n}=0, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n} \prod_{\substack{\{\gamma_{j_{n-1}}=1,\\ \gamma_{j_n}=1, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n} \Big/$$

$$\prod_{\substack{\{\gamma_{j_{n-1}}=1,\\ \gamma_{j_n}=0, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n} \prod_{\substack{\{\gamma_{j_{n-1}}=0,\\ \gamma_{j_n}=1, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n} \prod_{\substack{\{\gamma_{j_{n-1}}=0,\\ \gamma_{j_n}=0, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n} \prod_{\substack{\{\gamma_{j_{n-1}}=1,\\ \gamma_{j_n}=1, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T,j_1\ldots j_n}$$

$$= \prod_{\substack{\{\gamma_{j_{n-1}}=1, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(1-p_1)(1-p_2) \prod_{\substack{\{\gamma_{j_{n-1}}=0, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(1-p_1)p_2$$

$$\prod_{\substack{\{\gamma_{j_{n-1}}=0, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(p_1+(1-p_1)(1-p_2)) \prod_{\substack{\{\gamma_{j_{n-1}}=1, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(p_1+(1-p_1)p_2) \Big/$$

$$\prod_{\substack{\{\gamma_{j_{n-1}}=1, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(1-p_1)(1-p_2) \prod_{\substack{\{\gamma_{j_{n-1}}=0, \text{ even \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(1-p_1)p_2$$

$$\prod_{\substack{\{\gamma_{j_{n-1}}=0, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(p_1+(1-p_1)(1-p_2)) \prod_{\substack{\{\gamma_{j_{n-1}}=1, \text{ odd \# of 0's}\\ \text{in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}}} p^*_{T-1,j_1\ldots j_{n-1}}(p_1+(1-p_1)p_2)$$

$$= \prod_{\{\text{odd \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} (1-p_1)(1-p_2) \Big/ \prod_{\{\text{even \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} (1-p_1)(1-p_2)$$

$$\prod_{\{\text{odd \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} (1-p_1)p_2 \Big/ \prod_{\{\text{even \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} (1-p_1)p_2$$

$$\prod_{\{\text{even \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} p_1+(1-p_1)(1-p_2) \Big/ \prod_{\{\text{odd \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} p_1+(1-p_1)(1-p_2)$$

$$\prod_{\{\text{even \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} p_1+(1-p_1)p_2 \Big/ \prod_{\{\text{odd \# of 0's in } \gamma_{j_1},\ldots,\gamma_{j_{n-2}}\}} p_1+(1-p_1)p_2,$$

which equals $\left(p_1+p_2(1-p_2)(1-p_1)^2\right)/p_2(1-p_2)(1-p_1)^2$ if $n = 2$ and 1 if $n > 2$.

Therefore we have for $T = 3$,

$$\exp(f_3^{123}) = 1, \quad \exp(f_3^{23}) = \frac{p_1 + p_2(1-p_2)(1-p_1)^2}{p_2(1-p_2)(1-p_1)^2}.$$

Therefore the NDARMA(1) model with parameter $(p1, p2)$ has the multivariate Bernoulli probability mass function where

$$\exp(f_T^1) = \exp(f_T^T) = \frac{p_2(1-p_1)}{p_1 + (1-p_2)(1-p_1)},$$

$$\exp(f_T^{j_1}) = \frac{p_2(1-p_2)(1-p_1)^2}{(p_1 + (1-p_2)(1-p_1))^2}, \quad j_1 = 2, \ldots, T-1,$$

$$\exp(f_T^{j_1 j_2}) = \frac{p_1 + p_2(1-p_2)(1-p_1)^2}{p_2(1-p_2)(1-p_1)^2}, \quad j_2 = j_1 + 1, j_1 = 1, \ldots, T-1,$$

and the rest $f_T^{j_1 j_2 \ldots j_n}, n = 2, \ldots, T$ are all 0 such that $\exp(f_T^{j_1 j_2 \ldots j_n}) = 1$.

Lastly note that f_T^1 and f_T^T are θ, $f_T^{j_1}$ for $j_1 = 2, \ldots, T-1$ are θ^* and $f_T^{j_1 j_2}$, when j_1 and j_2 are two neighboring integers, are identical to κ. This concludes the proof. □

Markov Chain Monte Carlo algorithm

We report the Gibbs sampler used to sample from the posterior distribution of the time-varying tensor vector auto-regressive model under our prior specification. For notational simplicity, we summarize $\mathbf{y}_1, \ldots, \mathbf{y}_T$ with Y, and we use the capital letters Φ, Λ, V and Z to denote the vectors containing the elements ϕ_h, $\lambda_{1,h}$,$\lambda_{2,h}$,$v_{h,k}$,$z_{h,k}$ for varying values of $h = 1, \ldots, H$, and $k = 1, \ldots, P$, respectively. We further employ the capital letter A for the vector containing the values $\alpha_{1,h}$,$\alpha_{2,h}$,$\alpha_{3,h}$ for varying $h = 1, \ldots, H$ and the capital letter Γ for the vector of $\gamma_{h,t}$'s, $h = 1, \ldots, H$,$t = P+1, \ldots, T$. Likewise, W is used to denote the collection of $w_{1,h,k}$, $w_{2,h,k}$,$k = 1, \ldots, N$, and $w_{3,h,k}$, $k = 1, \ldots, P$.

1. Sample from the full conditional

$$\mathrm{pr}(\alpha, \Phi, \tau \mid A, W) = \mathrm{pr}(\alpha \mid A, W)\mathrm{pr}(\Phi \mid \alpha, A, W)\mathrm{pr}(\tau \mid \alpha, \Phi, A, W)$$

as follows:

   (a) sample $\mathrm{pr}(\alpha \mid A, W)$ using a griddy-Gibbs: for each α in a set of even spaced grid values in interval $(H^{-3}, H^{-0.1})$, draw M samples of ϕ and τ as in b) and c). Sample α from the grid values with probability proportional to the average score $\mathrm{pr}(A \mid \phi, \tau, W)\mathrm{pr}(\phi, \tau \mid \alpha)$ over M samples.

   (b) sample $\psi_1, \ldots, \psi_H$ independently from the generalized inverse Gaussian distribution $\mathrm{giG}(\alpha - N - P/2, 2b_\tau, 2C_h)$ where $C_h = \sum_{j=1}^{3} \alpha'_{j,h} W_{j,h}^{-1} \alpha_{j,h}$; then, normalize to have $\phi_1, \ldots, \phi_H$

   (c) sample τ from $\mathrm{giG}(a_\tau - H(N - P/2), 2b_\tau, 2\sum_{h=1}^{H} C_h/\phi_h)$

2. Sample from the full conditional

$$\mathrm{pr}(\Lambda, V, Z, W \mid A, \Phi, \tau)$$

as follows:

   (a) sample $\lambda_{1,h}, \lambda_{2,h}$ from $\mathrm{Ga}(a_\lambda + N, b_\lambda + \|\alpha_{1,h}\|_1/\sqrt{\phi_h \tau})$ and $\mathrm{Ga}(a_\lambda + N, b_\lambda + \|\alpha_{2,h}\|_1/\sqrt{\phi_h \tau})$, respectively

   (b) sample $W_{1,h,k}, W_{2,h,k}$ from $\mathrm{giG}(1/2, \lambda_{1,h}^2, \alpha_{1,h,k}^2/\phi_h\tau)$ and $\mathrm{giG}(1/2, \lambda_{2,h}^2, \alpha_{2,h,k}^2/\phi_h\tau)$ respectively

   (c) sample $v_{h,k}$ from Beta $\left(1 + \sum_{p=1}^{P} \mathbb{1}\left(z_{h,p} = k\right), \beta + \sum_{p=1}^{P} \mathbb{1}\left(z_{h,p} > k\right)\right)$

(d) sample $z_{h,k}$ from

$$\mathrm{pr}\left(z_{h,k} = l \mid w_{h,1}, \ldots, w_{h,P}, \alpha_{3,h}\right) \propto \begin{cases} w_{h,l}\mathcal{N}\left(\alpha_{3,h,k}; 0, \tau\phi_h W_\infty\right), & l \leq k \\ w_{h,l}t_{2a_w}\left(\alpha_{3,h,k}; 0, b_w\tau\phi_h/a_w\right), & l > k \end{cases}$$

where t_{2a_w} stands for the generalized Student's t-distribution with $2w$ degrees of freedom

(e) sample $W_{3,h,k}$ as follows

$$W_{3,h,k} \begin{cases} = W_\infty, & \text{if } z_{h,k} \leq k \\ \sim \mathrm{IG}\left(a_w + 1/2, b_w + \alpha_{3,h,k}^2/(2\tau\phi_h)\right), & \text{if } z_{h,k} > k \end{cases}$$

3. Sample from

$$\mathrm{pr}(A \mid W, \Phi, \tau, \sigma, \mathbf{y})$$

as follows:

(a) sample $\alpha_{1,h}$ from $\mathcal{N}(\mu_{1,h}, \Sigma_{1,h})$ with

$$\Sigma_{1,h} = \left(W_{1,h}^{-1}/\phi_h\tau + \sum_{t=P+1}^{T} D_{1,h,t}^2\Sigma^{-1}\right)^{-1}$$

and

$$\mu_{1,h} = \Sigma_{1,h}\left(\sum_{t=P+1}^{T} D_{1,h,t}\Sigma^{-1}\tilde{\mathbf{y}}_{1,t,h}\right)$$

where

$$D_{1,h,t} = \gamma_{h,t}(\alpha_{3,h} \otimes \alpha_{2,h})' \left[\mathbf{y}'_{t-1} \ldots \mathbf{y}'_{t-P}\right]'$$

and

$$\tilde{\mathbf{y}}_{1,t,h} = \mathbf{y}_t - \sum_{j=1,j\neq h}^{H} D_{1,t,h}\alpha_{1,j}$$

(b) sample $\alpha_{2,h}$ from $\mathcal{N}(\mu_{2,h}, \Sigma_{2,h})$ with

$$\Sigma_{2,h} = \left(W_{2,h}^{-1}/\phi_h \tau + \sum_{t=P+1}^{T} D'_{2,h,t} \Sigma^{-1} D_{2,h,t} \right)^{-1}$$

and

$$\mu_{2,h} = \Sigma_{2,h} \left(\sum_{t=P+1}^{T} D'_{2,h,t} \Sigma^{-1} \tilde{\mathbf{y}}_{2,t,h} \right)$$

where

$$D_{2,h,t} = \gamma_{h,t} (\alpha'_{3,h} \otimes \alpha_{2,h}) \left[\mathbf{y}_{t-1} \dots \mathbf{y}_{t-P} \right]'$$

and

$$\tilde{\mathbf{y}}_{2,t,h} = \mathbf{y}_t - \sum_{j=1, j \neq h}^{H} D_{2,h,t} \alpha_{2,j}$$

(c) sample $\alpha_{3,h}$ from $\mathcal{N}(\mu_{3,h}, \Sigma_{3,h})$ with

$$\Sigma_{3,h} = \left(W_{3,h}^{-1}/\phi_h \tau + \sum_{t=P+1}^{T} D'_{3,h,t} \Sigma^{-1} D_{3,h,t} \right)^{-1}$$

and

$$\mu_{3,h} = \Sigma_{3,h} \left(\sum_{t=P+1}^{T} D'_{3,h,t} \Sigma^{-1} \tilde{\mathbf{y}}_{3,t,h} \right)$$

where

$$D_{3,h,t} = \gamma_{h,t} \left[(\alpha_{1,h} \circ \alpha_{2,h}) \mathbf{y}_t, \dots, (\alpha_{1,h} \circ \alpha_{2,h}) \mathbf{y}_{t-P} \right]$$

and

$$\tilde{\mathbf{y}}_{3,t,h} = \mathbf{y}_t - \sum_{j=1, j \neq h}^{H} D_{3,j,t} \alpha_{3,j}$$

4. Sample from

$$\mathrm{pr}(\Gamma, \theta_h, \kappa_h \mid A, \sigma, Y)$$

as follows:

(a) sample $\gamma_{h,t}, t = P+1, \dots, T$ from

$$\text{pr}\left(\gamma_{h,P+1}, \dots, \gamma_{h,T} \mid \cdot\right) \propto \exp\left(\sum_{t=P+1}^{T} \left(\psi_{h,t} + \theta_{h,t}\right)\gamma_{h,t} + \sum_{t=P+1}^{T-1} \kappa_h \gamma_{h,t}\gamma_{h,t+1}\right)$$

with

$$\psi_{h,t} = \bar{\mathbf{y}}'_{h,t}\Sigma^{-1}\tilde{\mathbf{y}}_{h,t} - 1/2\bar{\mathbf{y}}'_{h,t}\Sigma^{-1}\bar{\mathbf{y}}_{h,t}$$

where

$$\bar{\mathbf{y}}_{h,t} = \alpha'_{3,h} \otimes \left(\alpha_{1,h} \circ \alpha_{2,h}\right)\left[\mathbf{y}'_{t-1} \dots \mathbf{y}'_{t-P}\right]'$$

and

$$\tilde{\mathbf{y}}_{h,t} = \mathbf{y}_t - \sum_{j=1, j\neq h}^{H} \gamma_{j,t}\bar{\mathbf{y}}_{h,j}$$

(b) sample θ_h and κ_h using the auxiliary variable method by Møller et al. (2006) and the Propp and Wilson (1996) perfect sampling algorithm.

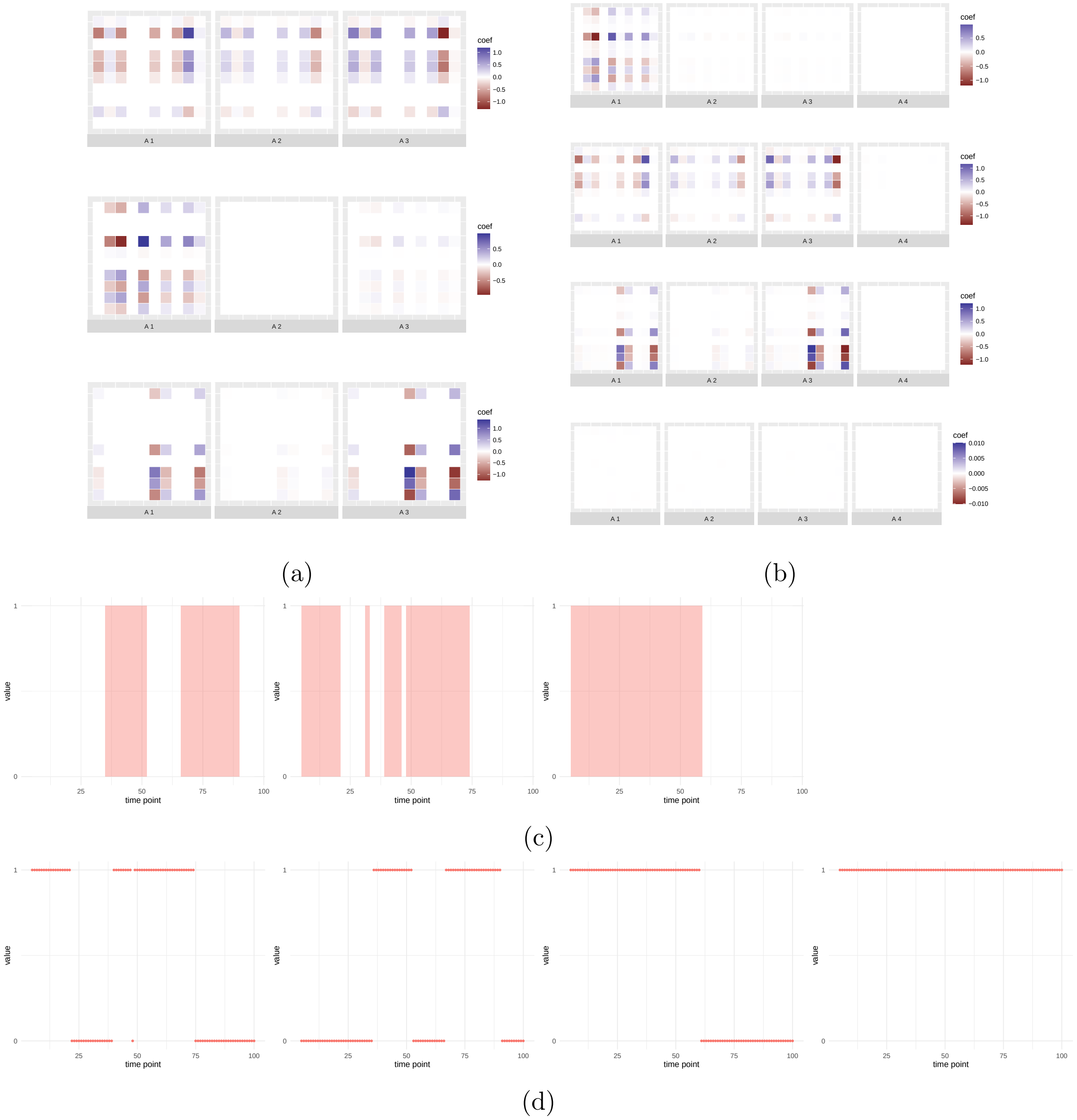
5. sample σ_n^2 from a $\text{IG}(a_\sigma + (T-P)/2, b_\sigma + \|\tilde{\mathbf{y}}_t\|_2/2)$ where

$$\tilde{\mathbf{y}}_t = \mathbf{y}_t - \sum_{h=1}^{H} \gamma_{h,t}\alpha'_{3,h} \otimes \left(\alpha_{1,h} \circ \alpha_{2,h}\right)\left[\mathbf{y}'_{t-1} \dots \mathbf{y}'_{t-P}\right]'$$

Some simulation results

We show in Figure 3.9 and Figure 3.10 results of 2 trials in the first simulation study as examples. Our method successfully identifies all three components in trial 12 and the estimated binary trajectories recover the true activation, while in trial 47, only one component is correctly estimated. Having in mind that there are still non-negligible components that our method fails to detect, we would like to offer some preliminary explanation. One hypothesis is when the magnitude and the sparsity of the VAR coefficient matrix or when the component barely activates during the process, it become difficult to identify because of the weak signal. In Figure 3.11, we show the boxplot of some metrics

$(\gamma_{h,P+1}, \ldots, \gamma_{h,T})$ and $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$ in two groups, one where the component is identified and the other not. From the boxplot, it can be seen that the sparsity of the sequence $(\gamma_{h,P+1}, \ldots, \gamma_{h,T})$ and the magnitude of the coefficient $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$, either in max norm or in Frobenius norm, are related to the identifiability of that component as expect. However, the two groups are not distinguishable if we look at the sparsity of the coefficient matrix. Additionally We can investigate the hypothesis taking into account both the sparsity of $(\gamma_{h,P+1}, \ldots, \gamma_{h,T})$ and the magnitude of $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$. In Figure 3.12, we can compute the line to separate the identified and the non-identified classification by the support vector machine (SVM). The green points (the non-identified group) mostly allocate in the left bottom area of the plane whereas the red points (the identified group) take the space where the signal is supposedly large. As a final remark, it is important to bare in mind that the above analysis only gives a rough idea why some components are missed in the estimation and the metrics that we use are not necessarily the real factors that cause the issue.



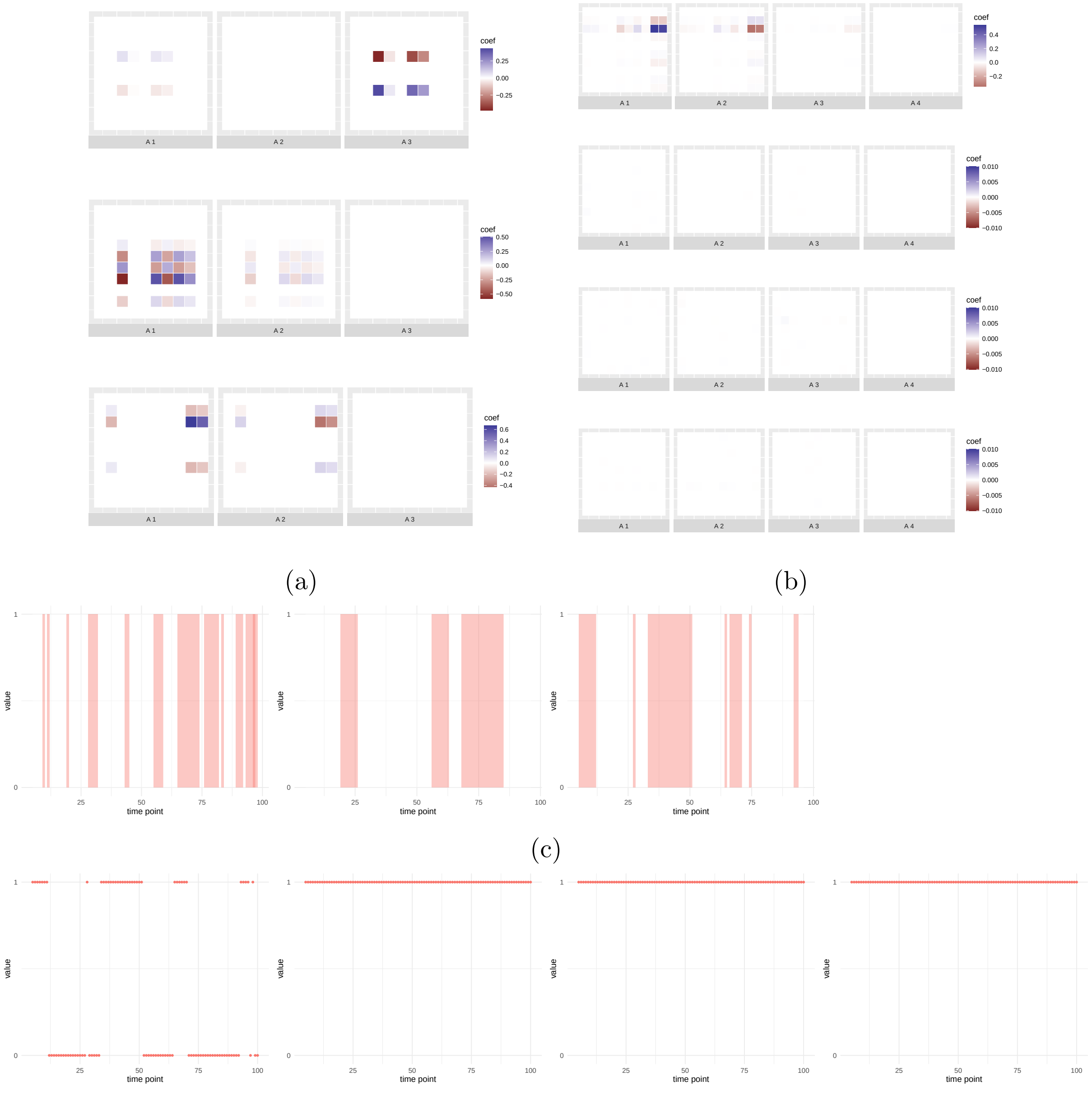
(a)

(b)

(c)

(d)

Figure 3.9: Results of trial 12 in the first simulation study: (1) the three true components; (2) the four estimated components; (3) true regions of activation associated with each component, where red indicates the component is active inside the time interval; (4) estimated activation trajectories corresponding to each component.



(a)

(b)

(c)

(d)

Figure 3.10: Results of trial 47 in the first simulation study: (1) the three true components; (2) the four estimated components; (3) true regions of activation associated with each component, where red indicates the component is active inside the time interval; (4) estimated activation trajectories corresponding to each component.

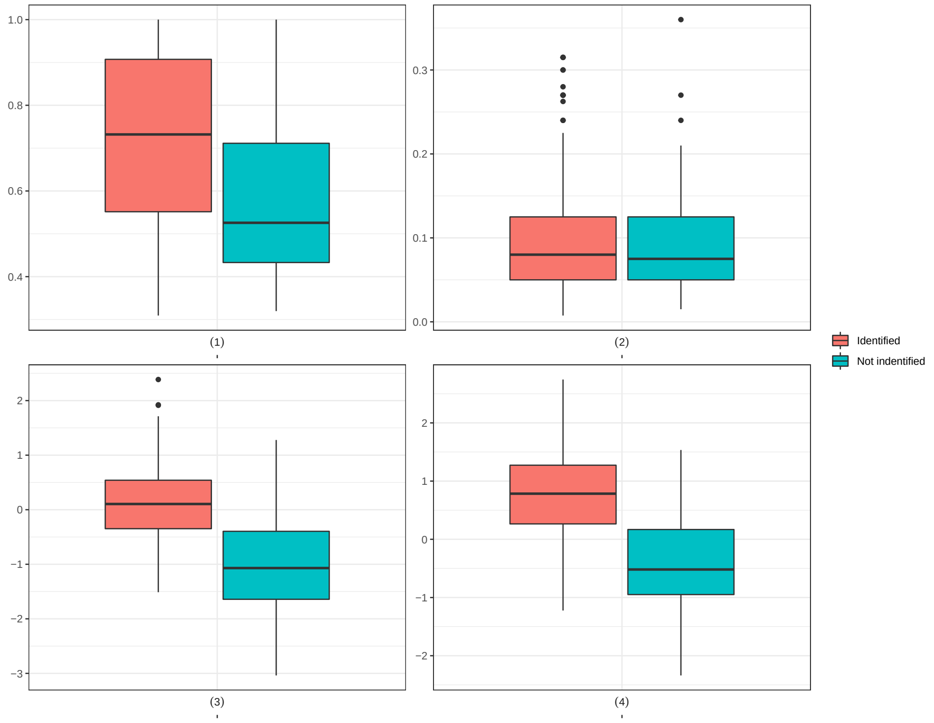


Figure 3.11: Boxplots of certain metrics of $(\gamma_{h,P+1}, \ldots, \gamma_{h,T})$ and $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$ grouped by whether the component is identified in posterior samples. Sparsity means the proportion of non zero elements.

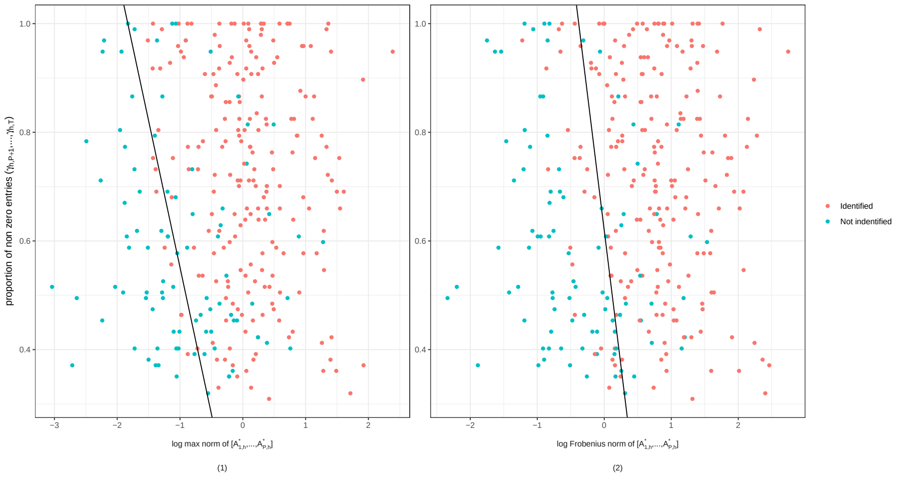


Figure 3.12: A simple illustration showing how the sparsity of $(\gamma_{h,P+1}, \ldots, \gamma_{h,T})$ and the norm of $\left[A^*_{1,h}, A^*_{2,h}, \ldots, A^*_{P,h}\right]$ classify identifiability. The black line is obtained using the support vector machine (SVM).

Chapter 4

Dynamic Shrinkage Priors in Time-Varying Vector Autoregressive Models

Abstract

The vector autoregressive (VAR) model is a popular choice for multivariate time series analysis, however it does not apply to non-stationary time series. A possible remedy is to extend the VAR model to time-varying VAR (TV-VAR) models to allow for changing coefficients. We address two statistical challenges when applying the TV-VAR model, the issue of high dimensionality and the modeling of temporal dependence. We exploit lower dimensional representation of the parameter space in TV-VAR models via tensor factorization to achieve effective dimension reduction. In addition, the temporal dependence of sparsity along time is modeled using two prior specification, the dynamic spike and slab (DSS prior and the dynamic shrinkage processes (DSP) priors. In many application, experiments are conducted multiple times under the same experimental conditions and it is ideal that a model be able

to pool information from multiple trials to make statistical inferences. Our proposed model introduces a hierarchical structure in the parameter space to account for data of this type. The method is validated through simulation studies and applied to local field potentials (LFPs) data when rats perform a hippocampus-dependent sequence-memory task.

4.1 Introduction

In many applications, observations are taken over time and the parameters governing the data generating process are assumed to vary with time. For example, it is well documented in economics that the role of financial indicators in predicting inflation changes in different stages of an economic cycle (Stock and Watson, 2007; Cogley et al., 2010). Similarly, in neuroscience, recent literature has discovered changes in brain connectivity in response to a series of stimuli in task-based experimental settings or because of inherent spontaneous fluctuations in resting state fMRI (Taghia et al., 2017; Warnick et al., 2018; Zarghami and Friston, 2020b). The need to study these phenomena motivates the choice of models with time-varying parameters that generalizes from their static counterparts to allow for more flexibility in modeling the dynamic patterns embedded in observed time series data. Simple examples include dynamic regression as an extension of ordinary regression (Petris et al., 2009; Prado and West, 2010; Pankratz, 2012) and the time-varying VAR (TV-VAR) model in terms of generalizing the VAR model (Primiceri, 2005; Nakajima et al., 2011). Another motivation to consider time-varying parameter models lies on the fact that any non-linear model can be approximated by a time-varying parameter linear model (Granger, 2008). Therefore, relatively simpler time-varying parameter models avoid the use of sophisticated non-linear models, greatly mitigating technical complexity inherent to non-linear models.

In this chapter, we explore alternative approaches to model temporal dependence in the dynamic coefficient matrix of the time-varying vector autoregressive (TV-VAR) model. The dynamics proposed in Chapter 3 are modeled via a Markovian switching state characterization with the evolution being governed by a temporal Ising prior that ultimately assumes Markovian dependence structure. By construction, each of the tensor decomposition components can either be active or non-active at any time point. Consequently, the coefficient matrix may transit abruptly from one state to another and the size of the state space is exponential in the tensor decomposition rank. Instead it seems interesting to consider that the system has smooth transitions from one state of mind to another state, which calls for smooth temporal evolution of the time-varying parameters. The general class of state space model offers advantages in terms of interpretation and well-established inference procedures. Here we adopt this view and regard a TV-VAR as a state space model where the parameters are latent states. However, without any further dimension reduction of the TV-VAR model, the number of parameters is quadratic in the dimension of the time series. The sheer number of parameters to be estimated leads to low statistical and computational efficiency. For this reason, in recent years many dimension reduction techniques for TV-VAR models have been proposed in the literature. For example, Chan et al. (2020) use a low-rank approximation of the covariance matrix of the TV-VAR coefficients to effectively express them as a linear combination of lower-dimensional factors. Our approach boils down to applying the tensor decomposition to exploit the lower dimensional structure of the dynamic coefficient matrix (Wang et al., 2021; Zhang et al., 2021). We start from a state space framework, but will specify the state evolution on the tensor margins obtained as a result of the factorization.

Both frequentists and Bayesian methods have been proposed for TV-VAR models (Cogley and Sargent, 2005; Primiceri, 2005; Koop and Korobilis, 2013; Prieto et al., 2016; Samdin et al., 2016; Casas et al., 2017; Kapetanios et al., 2019). In particular, Bayesian inference for TV-VAR models usually specifies the evolution of coefficients as a first order Markov chain. When assuming Markovian transition, such model specification focuses on capturing the temporal dependence between two parameters consecutively in time while neglecting possible relationships of shrinkage effects along time. In fact, one may want to consider such dependences, i.e. to allow the shrinkage at a certain time point to depend on the shrinkage level at the previous time point. Rephrasing in terms of variable selection, this is to say that parameter values evolve smoothly over time and their importance exhibits some type of dependence over time. Several investigators have developed careful designs of dynamic priors on parameters in the context of Bayesian variable selection for time series. For example, Kalli and Griffin (2014) propose a Normal-Gamma autoregressive (NGAR) process prior to account for both time-varying regression coefficients and time-varying sparsity. Similar effect of dynamic variable selection can be achieved by the dynamic spike and slab (DSS) prior in Rockova and McAlinn (2021). A minute-by-minute Twitter central processor unit (CPU) usage data set motivates Kowal et al. (2019) to introduce the framework that models dependence between prior scale parameters. Different from above dynamic shrinkage priors, Irie (2019) takes on the issue by considering penalizing functions and extends the fused LASSO penalization to Bayesian dynamic fused LASSO. The aforementioned approaches have their merits but they are all initially designed for dynamic regression. A novel contribution in this chapter is to investigate the use of time-varying shrinkage priors in TV-VAR models, more specifically the TV-VAR model for multiple homogeneous trials under same experiment condition. The Bayesian hierarchical specification allows borrowing information across trials.

The rest of the chapter is organized as follows. In Section 4.2 we describe the TV-VAR model for multiple trials and its representation as a conditionally Gaussian state space model. The suggested tensor decomposition of the stacked time-varying coefficients is also formulated to define the time-varying tensor VAR (TVT-VAR) model. Section 4.3 elaborates the application of two specific dynamic shrinkage priors, the DSS and the DSP prior, in the context of TVT-VAR models. Section 4.4 describes the Markov chain Monte Carlo (MCMC) algorithm to draw posterior samples under the two prior specification. Results of two simulation studies are displayed in Section 4.5 to assess the performance of the proposed method. Section 4.6 applies the TVT-VAR model with dynamic shrinkage priors to rats' local field potentials (LFPs) data when performing a hippocampus-dependent sequence-memory task.

4.2 Time-varying Vector Autoregressive Models

Let $\mathbf{y}_t^s$ be an N-dimensional vector of observations of the same subject at time points $t = 1, \dots, T$ in trials $s = 1, \dots, S$ replicated under the same experiment condition. The TV-VAR model of order P assumes that $\mathbf{y}_t^s$ is a linear combination of the P lagged signals $\mathbf{y}_{t-1}^s, \dots, \mathbf{y}_{t-P}^s$ plus an independent noise $\boldsymbol{\epsilon}_t^s \in \mathbb{R}^N$,

$$\mathbf{y}_t^s = \begin{bmatrix} A_{1,t}^s, A_{2,t}^s, \dots, A_{P,t}^s \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1}^s \\ \vdots \\ \mathbf{y}_{t-P}^s \end{bmatrix} + \boldsymbol{\epsilon}_t^s, \tag{4.1}$$

where $\boldsymbol{\epsilon}_t^s \sim \mathcal{N}(0, \Sigma_t^s)$ and the linear coefficients $A_{p,t}^s$, $p = 1, \dots, P$ are trial specific $N \times N$ matrices, assumed to vary across time t. Here, we assume that the covariance matrix Σ_t^s of the error term $\boldsymbol{\epsilon}_t^s$ does not depend on time t, i.e. $\Sigma_t^s = \Sigma^s$ as we do not consider stochastic volatility for simplicity.

The model is completed with the specification of the temporal evolution of the $A_{p,t}^s$ matrices. In the literature, this is usually achieved by formulating the TV-VAR as a linear and Gaussian state-space model or dynamic linear model (DLM) (Prado and West, 2010). However, the autocorrelation structure of the time series $(\mathbf{y}_t^s)$ imposes to 'force' the DLM specification, usually by setting the covariance matrix of the observation error to zero. Here instead we move from a representation of the TV-VAR model as a *conditionally Gaussian state space model* (Lipster and Shiryayev, 1972; Liptser and Shiryaev, 2013) which extends DLMs, in particular by allowing the system matrix in the observation equation to depend on past values of the observations (Lipster and Shiryayev, 1972). Treating $A^s_{1,t}, \dots A^s_{P,t}$ as latent state variables, the observation equation in the conditionally Gaussian state space representation can be written as

$$\mathbf{y}^s_t = F^s_t \begin{bmatrix} vec({A^s_{1,t}}^\top) \\ \vdots \\ vec({A^s_{P,t}}^\top) \end{bmatrix} + \boldsymbol{\epsilon}^s_t, \quad \boldsymbol{\epsilon}^s_t \sim \mathcal{N}(0, \Sigma^s), \tag{4.2}$$

where $F^s_t = \begin{bmatrix} \mathbf{I}_N \otimes {\mathbf{y}^s_{t-1}}^\top, \dots, \mathbf{I}_N \otimes {\mathbf{y}^s_{t-P}}^\top \end{bmatrix}$ is a sparse $N \times N^2P$ matrix. The vector of trial specific state variables $\begin{bmatrix} vec({A^s_{1,t}}^\top)^\top, \dots, vec({A^s_{P,t}}^\top)^\top \end{bmatrix}^\top$ has length N^2P.

This model refers to a single trial s. Since the S trials are conducted on the same single subject, it is reasonable to pool them together assuming exchangeability across trials, and thus allow borrowing of strength across trials in learning the underlying subject specific coefficients. We achieve the goal by modeling the probabilistic dependence across trials through a hierarchical model that assumes that the trial-specific coefficients are conditionally independent and Gaussian given the overall mean vector $\begin{bmatrix} vec({A_{1,t}}^\top)^\top, \dots, vec({A_{P,t}}^\top)^\top \end{bmatrix}^\top$ and the diagonal covariance matrix Λ

$$\begin{bmatrix} vec({A^s_{1,t}}^\top) \\ \vdots \\ vec({A^s_{P,t}}^\top) \end{bmatrix} \Bigg| \begin{bmatrix} vec({A_{1,t}}^\top) \\ \vdots \\ vec({A_{P,t}}^\top) \end{bmatrix}, \Lambda \overset{indep}{\sim} \mathcal{N}\left(\begin{bmatrix} vec({A_{1,t}}^\top) \\ \vdots \\ vec({A_{P,t}}^\top) \end{bmatrix}, \Lambda \right) \quad s = 1, \dots, S. \tag{4.3}$$

Consequently, the temporal dependence of trial specific coefficients can be specified by modeling the dynamics of the vector of the subject's means. When the evolution of latent variables $\begin{bmatrix} vec({A_{1,t}}^\top)^\top, \dots, vec({A_{P,t}}^\top)^\top \end{bmatrix}^\top$ is both linear and Gaussian, we obtain a conditionally Gaussian state space representation of the TV-VAR model for multiple trials after marginalizing out $\begin{bmatrix} vec({A^s_{1,t}}^\top)^\top, \dots, vec({A^s_{P,t}}^\top)^\top \end{bmatrix}^\top$ and concatenating $\mathbf{y}^s_t$ to form $\mathbf{y}_t =$

$((\mathbf{y}_t^1)^\top, \dots, (\mathbf{y}_t^S)^\top)^\top$.

Although this modeling choice is straightforward to interpret and may be effectively applied for small time series dimension N, the size of the latent space grows quadratically with N, creating statistical and computational challenges to inference especially when N becomes large. To address this dimensionality issue, we propose to first stack $\left[A_{1,t}^s, A_{2,t}^s, \dots, A_{P,t}^s\right]$ into a three-way tensor $\mathcal{A}_t^s$ of size $N \times N \times P$ and then apply a PARAFAC decomposition to achieve an increased reduction in the number of estimands. Tensors decomposition has been widely applied to multidimensional array data (Zhou et al., 2013; Guhaniyogi et al., 2017). In general, a $q_1 \times q_2 \times \cdots \times q_M$ tensor $\mathcal{A}$ is said to admit a rank-R PARAFAC decomposition if R is the smallest integer such that $\mathcal{A}$ can be written as

$$\mathcal{A} = \sum_{r=1}^{R} \alpha_{1,r} \circ \alpha_{2,r} \circ \cdots \circ \alpha_{M,r},$$

where $\circ$ indicates the vector outer product and $\alpha_{m,r} \in \mathbb{R}^{q_m}, m = 1, \dots, M$ are the *tensor margins* of each mode. In our case the transformed three-way tensor $\mathcal{A}_t^s$ is assumed to consist of H lower rank components such that

$$\mathcal{A}_t^s = \sum_{h=1}^{H} \boldsymbol{\alpha}_{1,t,h}^s \circ \boldsymbol{\alpha}_{2,t,h}^s \circ \boldsymbol{\alpha}_{3,t,h}^s$$

with $\boldsymbol{\alpha}_{1,t,h}^s, \boldsymbol{\alpha}_{2,t,h}^s \in \mathbb{R}^N$ and $\boldsymbol{\alpha}_{3,t,h}^s \in \mathbb{R}^P$. Now the original mean coefficient matrix can be recovered by rearranging the modes of the tensor decomposition as follow,

$$\left[A_{1,t}^s, A_{2,t}^s, \dots, A_{P,t}^s\right] = \sum_{h=1}^{H} {\boldsymbol{\alpha}_{3,t,h}^s}^\top \otimes \left(\boldsymbol{\alpha}_{1,t,h}^s \circ \boldsymbol{\alpha}_{2,t,h}^s\right). \tag{4.4}$$

Note that for $p = 1, \dots, P$, each matrix $A_{p,t}^s = \sum_{h=1}^{H} \alpha_{3,p,t,h}^s \cdot (\boldsymbol{\alpha}_{1,t,h}^s \circ \boldsymbol{\alpha}_{2,t,h}^s)$ where $\alpha_{3,p,t,h}^s$ denotes the p-th entry of vector $\boldsymbol{\alpha}_{3,t,h}^s$, reflecting a sequence of constraints on the coefficient matrix: the element-by-element ratio between $A_{1,t}^s$ and $A_{2,t}^s$ is proportional to the ratio between the first two entries of $\boldsymbol{\alpha}_{3,t,h}$, and similarly for subsequent lags. In the same spirit as before, we assume that the tensor $\mathcal{A}_t$ stacked from the subject specific coefficient matrix $\Big[A_{1,t}, A_{2,t}, \dots, A_{P,t}\Big]$ has the same PARAFAC lower dimensional structure

$$\mathcal{A}_t = \sum_{h=1}^{H} \boldsymbol{\alpha}_{1,t,h} \circ \boldsymbol{\alpha}_{2,t,h} \circ \boldsymbol{\alpha}_{3,t,h}.$$

Thus, the model is now parametrized in the tensor margins $\boldsymbol{\alpha}^s_{1,t,h}, \boldsymbol{\alpha}^s_{2,t,h}, \boldsymbol{\alpha}^s_{3,t,h}$. Therefore, it is natural to induce probabilistic dependence across trials through the joint probability law of the tensor margins, rather than of the original coefficient matrices $A^s_{p,t}$. We thus replace the modeling assumptions (4.3) by a new hierarchical model for the trial-specific tensor margins, that are assumed to be conditionally independent and normally distributed according to

$$\boldsymbol{\alpha}^s_{f,t,h} \mid \boldsymbol{\alpha}_{f,t,h}, \Lambda_{f,h} \overset{indep}{\sim} \mathcal{N}(\boldsymbol{\alpha}_{f,t,h}, \Lambda_{f,h}), \quad f = 1, 2, 3 \tag{4.5}$$

where $\Lambda_{1,h}, \Lambda_{2,h}$ are $N \times N$ diagonal matrices and $\Lambda_{3,h}$ is $P \times P$ diagonal matrix. The advantage of (4.5) over (4.3) is to effectively downsize the parameters in the initial TV-VAR model formulation from the original $(T-P)N^2P$ to $H(T-P)(2N+P)$ in each trial. Equations (4.1) (4.4) and (4.5) together define what we name the *Bayesian hierarchical time-varying tensor VAR* (BHTVT-VAR) model for multiple trials.

When tensor margins $\boldsymbol{\alpha}_{1,t,h}, \boldsymbol{\alpha}_{2,t,h}, \boldsymbol{\alpha}_{3,t,h}$ are regarded as latent states and proper temporal dependence is defined on them, one could in principle work out the induced dynamics on the original $A^s_{p,t}$ coefficients. However, this is usually not manageable, even in the case of simple linear Gaussian dependence in tensor margins. Remedy to this is to leverage the fact that the joint distribution can be uniquely specified given the set of complete conditional distributions under some compatibility conditions (Besag, 1974; Arnold and Press, 1989). For any fixed $f = 1, 2, 3$ and $h = 1, \dots, H$, conditionally on $\boldsymbol{\alpha}^s_{g,t,h}, g = 1, 2, 3, g \neq f$ and $\boldsymbol{\alpha}^s_{1,t,k}, \boldsymbol{\alpha}^s_{2,t,k}, \boldsymbol{\alpha}^s_{3,t,k}, k = 1, \dots, H, k \neq h$, and the P observations lags, the observation equation concerning $\boldsymbol{\alpha}^s_{f,t,h}$, obtained from (4.1) by subtracting $\sum_{k\neq h}^H \boldsymbol{\alpha}^s_{3,t,k}{}^\top \otimes \left(\boldsymbol{\alpha}^s_{1,t,k} \circ \boldsymbol{\alpha}^s_{2,t,k}\right)$ from its both sides, can be expressed as

$$\tilde{\mathbf{y}}^s_{t,h} = F^s_{f,t,h}\boldsymbol{\alpha}^s_{f,t,h} + \boldsymbol{\epsilon}^s_t, \tag{4.6}$$

where

$$\tilde{\mathbf{y}}^s_{t,h} = \mathbf{y}^s_t - \sum_{k\neq h}^{H} \boldsymbol{\alpha}^s_{3,t,k}{}^\top \otimes \left(\boldsymbol{\alpha}^s_{1,t,k} \circ \boldsymbol{\alpha}^s_{2,t,k}\right) \begin{bmatrix} \mathbf{y}^s_{t-1} \\ \vdots \\ \mathbf{y}^s_{t-P} \end{bmatrix}.$$

$F^s_{f,t,h}$ is equal to $F^s_{1,t,h} = (\boldsymbol{\alpha}^s_{3,t,h} \otimes \boldsymbol{\alpha}^s_{2,t,h})^\top \cdot (\mathbf{y}^s_{t-1}{}^\top, \ldots, \mathbf{y}^s_{t-P}{}^\top)^\top \cdot \mathbf{I}_N$ for $f = 1$, $F^s_{2,t,h} = (\boldsymbol{\alpha}^s_{3,t,h} \otimes \boldsymbol{\alpha}^s_{1,t,h})(\mathbf{y}^s_{t-1}, \ldots, \mathbf{y}^s_{t-P})^\top$ for $f = 2$, and $F^s_{3,t,h} = (\boldsymbol{\alpha}^s_{1,t,h} \cdot \boldsymbol{\alpha}^s_{2,t,h})^\top(\mathbf{y}^s_{t-1}, \ldots, \mathbf{y}^s_{t-P})$ for $f = 3$. The conditional specification (4.6) fully characterizes the BHTVT-VAR model for multiple trials. When combined with normality assumption in (4.5) and linear Gaussian state evolution of $\boldsymbol{\alpha}_{1,t,h}, \boldsymbol{\alpha}_{2,t,h}, \boldsymbol{\alpha}_{3,t,h}$, the BHTVT-VAR model can be seen as consisting of multiple conditionally Gaussian state space models for each $\boldsymbol{\alpha}_{f,t,h}$ after marginalizing out $\boldsymbol{\alpha}^s_{f,t,h}$ and concatenating $\tilde{\mathbf{y}}^s_{t,h}$ over s (see Section 4.4), providing a new angle to look into the model. The full conditional distributions are also essential in conducting posterior inferences.

In the next section, we describe the dynamic shrinkage priors on mean vectors $\boldsymbol{\alpha}_{1,t,h}$, $\boldsymbol{\alpha}_{2,t,h}, \boldsymbol{\alpha}_{3,t,h}$ that induce the desired linear Gaussian dynamics (or linear Gaussian dynamics conditioning on some hyperparameters).

4.3 Dynamic Shrinkage Priors

In order to model the temporal dependence of the tensor margins $\boldsymbol{\alpha}_{1,t,h}$, $\boldsymbol{\alpha}_{2,t,h}$, and $\boldsymbol{\alpha}_{3,t,h}$, we investigate the use of two recently proposed prior models for Bayesian variable selection in time-varying regression. We seek priors that enable modeling the parameters' dynamics and also enforce sparsity of the tensor decomposition over time. More specifically, we focus on the dynamic spike and slab (DSS) prior recently proposed by Rockova and McAlinn (2021) and the dynamic shrinkage process (DSP) prior of Kowal et al. (2019), and discuss their application in the BHTVT-VAR framework presented in the previous Sections. Before going into the details of the two prior choices, we mention that the priors for the diagonal elements of the covariance matrices Σ^s, $s = 1, \dots, S$ and $\Lambda_{f,h}$, $f = 1, 2, 3, h = 1, \dots, H$, are an $\mathrm{IG}(a_\sigma, b_\sigma)$ and $\mathrm{IG}(a_\lambda, b_\lambda)$, respectively, where $\mathrm{IG}(\cdot, \cdot)$ denotes the inverse gamma distribution.

4.3.1 DSS process priors in BHTVT-VAR models

We first consider the DSS prior by Rockova and McAlinn (2021), which has been proposed for the problem of variable selection in a dynamic linear model setting, where the set of active predictors is allowed to evolve over time. We apply the prior to modeling of the three tensor margins $\boldsymbol{\alpha}_{1,t,h}$, $\boldsymbol{\alpha}_{2,t,h}$ and $\boldsymbol{\alpha}_{3,t,h}$. Note that differently than in Rockova and McAlinn (2021), where the time-varying regression parameters are univariate, here we are interested in the activation of vectors in $\mathbb{R}^n$. More specifically, for $f = 1, 2, 3$, $h = 1, \dots, H$, and $t = 1, \dots, T$, we consider the conditional specification,

$$\pi(\boldsymbol{\alpha}_{f,t,h} | \boldsymbol{\gamma}_{f,t,h}, \boldsymbol{\alpha}_{f,t-1,h}) = (1 - \boldsymbol{\gamma}_{f,t,h}) \circ \psi_0(\boldsymbol{\alpha}_{f,t,h} | K_{f,0,h}) + \boldsymbol{\gamma}_{f,t,h} \circ \psi_1(\boldsymbol{\alpha}_{f,t,h} | \boldsymbol{\mu}_{f,t,h}, K_{f,1,h}),$$

where $\circ$ denotes the Hadamard (element-wise) product. This is a mixture of two components: a multivariate *spike* density, $\psi_0(\boldsymbol{\alpha}_{f,t,h} | K_{f,0,h})$, to capture irrelevant coefficients and a multivariate *slab* density $\psi_1(\boldsymbol{\alpha}_{f,t,h} | \boldsymbol{\mu}_{f,t,h}, K_{f,1,h})$ for the active coefficients, where $K_{f,0,h}$ and $K_{f,1,h}$ indicate $N \times N$ ($f = 1, 2$) or $P \times P$ ($f = 3$) variance-covariance matrices. The $\boldsymbol{\gamma}_{f,t,h}$ is a vector of latent binary indicators of activation, with probability $\mathrm{pr}(\boldsymbol{\gamma}_{f,t,h} = 1 | \boldsymbol{\alpha}_{f,t-1,h}) = \boldsymbol{\theta}_{f,t,h}$, i.e. $\boldsymbol{\gamma}_{f,t,h} | \boldsymbol{\alpha}_{f,t-1,h} \sim \mathrm{Bern}(\boldsymbol{\theta}_{f,t,h})$.

For simplicity, in the following we assume $K_{f,0,h} = \kappa_{f,0,h} \times \mathbf{I}_{r_f}$ and $K_{f,1,h} = \kappa_{f,1,h} \times \mathbf{I}_{r_f}$, where $r_f = N$ for $f = 1, 2$ and $r_f = P$ for $f = P$, i.e. the elements of the tensor margins

$\boldsymbol{\alpha}_{f,t,h}$, $f = 1,2,3$, are assumed as independent. Under this simplification, we denote $\psi_0(\boldsymbol{\alpha}_{f,t,h} \mid K_{f,0,h})$ by $\psi_0(\boldsymbol{\alpha}_{f,t,h} \mid \kappa_{f,0,h})$.

Possible choices of the spike density $\psi_0(\boldsymbol{\alpha}_{f,t,h}|\kappa_{f,0,h})$ include Gaussian or Laplace densities centered around 0 with small variances $\kappa_{f,0,h}$, corresponding to a Ridge or LASSO regression, respectively. In the following, we use the Gaussian spike as it leads to simpler posterior computations in our framework.

The *slab* component has an auto-regressive specification since the mean $\boldsymbol{\mu}_{f,t,h} = \boldsymbol{\phi}_{f,0,h} + \boldsymbol{\phi}_{f,1,h} \circ (\boldsymbol{\alpha}_{f,t-1,h} - \boldsymbol{\phi}_{f,0,h})$, where the parameter vector $\boldsymbol{\phi}_{f,1,h}$ has elements $|\boldsymbol{\phi}^i_{f,1,h}| < 1$, $i = 1, \dots, r_f$, $f = 1,2,3$. The common variance is assumed large, $\kappa_{f,1,h} \gg \kappa_{f,0,h}$, to allow capturing large values of the active coefficients. Then, all the elements of the tensor margin vectors, $\boldsymbol{\alpha}_{f,t,h}{}'s$, $t = P+1, \dots, T$, are assumed to follow a stationary Gaussian AR(1) process with normal stationary distribution,

$$\psi_1^{ST}(\boldsymbol{\alpha}_{f,t,h}|\kappa_{f,1,h},\, \boldsymbol{\phi}_{f,0,h},\, \boldsymbol{\phi}_{f,1,h}) = \psi_1\left(\boldsymbol{\alpha}_{f,t,h} \mid \boldsymbol{\phi}_{f,0,h},\, \kappa_{f,1,h} \cdot \mathrm{diag}\left\{1/\left(1 - \boldsymbol{\phi}^2_{f,1,h}\right)\right\}\right),$$

whose mean is $\boldsymbol{\phi}_{f,0,h}$ and covariance matrix is diagonal $\kappa_{f,1,h} \cdot \mathrm{diag}\left\{1/\left(1 - \boldsymbol{\phi}^2_{f,1,h}\right)\right\}$. When the choice of the spike and the slab distributions are normal, we easily obtain that the conditionally Gaussian distribution for $\boldsymbol{\alpha}_{f,t,h}$ given the indicator $\boldsymbol{\gamma}_{f,t,h}$ and the previous $\boldsymbol{\alpha}_{f,t-1,h}$ can be expressed as

$$\boldsymbol{\alpha}_{f,t,h} \mid \boldsymbol{\gamma}_{f,t,h}, \boldsymbol{\alpha}_{f,t-1,h} = \mathcal{N}\left(\boldsymbol{\gamma}_{f,t,h} \circ \boldsymbol{\mu}_{f,t,h},\;\; \kappa_{f,1,h} \cdot \mathrm{diag}\left\{\boldsymbol{\gamma}_{f,t,h}\right\} + \kappa_{f,0,h} \cdot \mathrm{diag}\left\{1 - \boldsymbol{\gamma}_{f,t,h}\right\}\right),$$

In terms of the state evolution equation, this implies that

$$\boldsymbol{\alpha}_{f,t,h} = G^{DSS}_{f,t,h} \boldsymbol{\alpha}_{f,t-1,h} + \boldsymbol{\omega}_{f,t,h}, \quad \boldsymbol{\omega}_{f,t,h} \sim \mathcal{N}(\mathbf{0},\, W^{DSS}_{f,t,h}) \tag{4.7}$$

where $G^{DSS}_{f,t,h} = \mathrm{diag}\{\boldsymbol{\gamma}_{f,t,h} \circ \boldsymbol{\phi}_{f,1,h}\}$ and $W^{DSS}_{f,t,h} = \mathrm{diag}\{\kappa_{f,1,h}\boldsymbol{\gamma}_{f,t,h} + \kappa_{f,0,h}(1 - \boldsymbol{\gamma}_{f,t,h})\}$ if

$\boldsymbol{\phi}_{f,0,h}$ is set to $\mathbf{0}$. Last but not least, the evolving inclusion probabilities are given by

$$\begin{aligned}\boldsymbol{\theta}_{f,t,h} &= \theta(\boldsymbol{\alpha}_{f,t-1,h}) \\ &= \frac{\Theta_{f,h}\psi_1^{ST}(\boldsymbol{\alpha}_{f,t-1,h}|\kappa_{f,1,h},\boldsymbol{\phi}_{f,0,h},\boldsymbol{\phi}_{f,1,h})}{\Theta_{f,h}\psi_1^{ST}(\boldsymbol{\alpha}_{f,t-1,h}|\kappa_{f,1,h},\boldsymbol{\phi}_{f,0,h},\boldsymbol{\phi}_{f,1,h}) + (1-\Theta_{f,h})\psi_0(\boldsymbol{\alpha}_{f,t-1,h}|\kappa_{f,0,h})},\end{aligned} \tag{4.8}$$

with larger value of hyperparameter $\Theta_{f,h}$ favors less shrinkage on $\boldsymbol{\alpha}_{f,t,h}$. The parameter $\boldsymbol{\theta}_{f,t,h}$ is essential in modeling the dynamic shrinkage effect. If $\boldsymbol{\alpha}_{f,t-1,h}$ is more likely to be from the stationary slab distribution – in particular, if $\Theta_{f,h}\psi_1^{ST}(\boldsymbol{\alpha}_{f,t-1,h}|\kappa_{f,1,h},\boldsymbol{\phi}_{f,0,h},\boldsymbol{\phi}_{f,1,h}) > (1-\Theta_{f,h})\psi_0(\boldsymbol{\alpha}_{f,t-1,h}|\kappa_{f,0,h})$ –, then the probability of activation $\boldsymbol{\theta}_{f,t,h}$ will be larger than 1/2 and the probability that $\gamma_{f,t,h} = 1$ will be high. Thus, $\boldsymbol{\alpha}_{f,t,h}$ will more likely to be drawn from the stationary slab distribution, same as $\boldsymbol{\alpha}_{f,t-1,h}$. To complete the state equation, the initial condition for $\boldsymbol{\alpha}_{f,0,h}$ is given by the stationary distribution of the DSS prior

$$\begin{aligned}\boldsymbol{\alpha}_{f,0,h} &\sim \pi^{ST}(\boldsymbol{\alpha}_{f,0,h} \mid \Theta_{f,h},\kappa_{f,1,h},\boldsymbol{\phi}_{f,0,h},\boldsymbol{\phi}_{f,1,h},\kappa_{f,0,h}) \\ &= \Theta_{f,h}\psi_1^{ST}(\boldsymbol{\alpha}_{f,0,h}|\kappa_{f,1,h},\boldsymbol{\phi}_{f,0,h},\boldsymbol{\phi}_{f,1,h}) + (1-\Theta_{f,h})\psi_0(\boldsymbol{\alpha}_{f,0,h}|\kappa_{f,0,h}).\end{aligned}$$

Finally, the vector of hyperparameters $\boldsymbol{\phi}_{f,1,h}$ contains the autoregressive parameters for each element, with values between -1 and 1. The assigned prior assumes that each element of the vector $\boldsymbol{\phi}_{f,1,h}$ is assigned the prior

$$\frac{1+\boldsymbol{\phi}_{f,1,h}^i}{2} \sim \text{Beta}(a_\phi, b_\phi), \quad i = 1,\ldots,r_f,$$

with $r_f = N$ if $f = 1,2$ or $r_f = P$ is $f = 3$.

4.3.2 DSP priors in BHTVT-VAR models

A second example of time-varying variable selection prior is represented by the DSP prior by Kowal et al. (2019). Here we assume that the tensor margins $\boldsymbol{\alpha}_{1,t,h}, \boldsymbol{\alpha}_{2,t,h}$ and $\boldsymbol{\alpha}_{3,t,h}$ evolve according to a random walk,

$$\boldsymbol{\alpha}_{f,t,h} = \boldsymbol{\alpha}_{f,t-1,h} + \boldsymbol{\omega}_{f,t,h}, \quad f = 1, 2, 3. \tag{4.9}$$

The DSP prior assumes that the error term $\boldsymbol{\omega}_{f,t,h}$ is normally distributed with mean $\mathbf{0}$ and variance-covariance matrix diag $\{\exp(\mathbf{g}_{f,t,h})\}$, where $r_f = N$ if $f = 1, 2$ and $r_f = P$ if $f = 3$. The vectors of parameters on the diagonal of the covariance matrix, $\mathbf{g}_{f,t,h}$, follow a AR(1) process

$$\mathbf{g}_{f,t,h} = \boldsymbol{\phi}_{f,0,h} + \boldsymbol{\phi}_{f,1,h} \circ (\mathbf{g}_{f,t-1,h} - \boldsymbol{\phi}_{f,0,h}) + \boldsymbol{\eta}_{f,t,h}, \tag{4.10}$$

where $\boldsymbol{\phi}_{f,1,h}$ is a r_f-dimensional vector of elements, with elements bounded in absolute value by 1. Each entry of $\boldsymbol{\eta}_{f,t,h}$ is identically and independently distributed as the Z-distribution $Z(a_\eta, b_\eta, 0, 1)$ characterized by the density function

$$f_Z(z) = \frac{e^{a_\eta z}}{B(a_\eta, b_\eta)(1 + e^z)^{a_\eta + b_\eta}},$$

where $B(a_\eta, b_\eta)$ indicated the beta function. The Z-distribution provides a general case for many important shrinkage priors, e..g the horseshoe prior (Carvalho et al., 2010) that corresponds to $a_\eta = b_\eta = 1/2$. When parameters $\boldsymbol{\phi}_{f,0,h} = \boldsymbol{\phi}_{f,1,h} = \mathbf{0}$, we retrieve exactly the static horseshoe prior in this case. The parameter $\boldsymbol{\phi}_{f,1,h}$ also controls the behavior of the "horseshoe" shape shrinkage weight which concentrates probability masses around 0 and 1 so true signal levels remain while noises are shrunk towards 0. The larger $0 < |\boldsymbol{\phi}^i_{f,1,h}| < 1$, $i = 1, \dots, r_f$, $f = 1, 2, 3$ is, the more entry-wise correlated $\mathbf{g}_{f,t-1,h}$ and $\mathbf{g}_{f,t,h}$ become and the denser the shrinkage parameter is distributed to both ends of the

$[0,1]$ interval. This is intuitive as larger entries $|\boldsymbol{\phi}^i_{f,1,h}|$ corresponds to greater persistence in shrinkage behavior so marginally states of aggressive shrinkage or little shrinkage are predominant (Kowal et al., 2019). The prior on $\boldsymbol{\phi}^i_{f,1,h}$ takes the same form

$$\frac{1+\boldsymbol{\phi}^i_{f,1,h}}{2} \sim \text{Beta}(a_\phi,\, b_\phi), \quad i = 1, \dots, r_f,$$

with $r_f = N$ if $f = 1, 2$ or $r_f = P$ is $f = 3$ as in the DSS prior whereas $\boldsymbol{\phi}^i_{f,0,h}$ is assumed to follow

$$\boldsymbol{\phi}^i_{f,0,h} \mid \sigma_\phi,\, \xi_{\phi,f,h} \sim \mathcal{N}(\log(\sigma^2_\phi/(T-P)),\, \xi^{-1}_{\phi,f,h}), \quad \xi_{\phi,f,h} \sim \text{PG}(1,0),$$

where PG(1,0) denotes Pólya-gamma random variables (Barndorff-Nielsen et al., 1982; Polson et al., 2013).

4.4 Posterior Inferences

To conduct inference on subject level dynamic coefficient matrices $\Big[A_{1,t}, A_{2,t}, \dots, A_{P,t}\Big]$ we use MCMC methods. More specifically, the posterior samples are drawn from their corresponding full conditional distributions in Gibbs sampler. One step common to both sampling algorithm under DSS and DSP prior specifications that facilitates posterior sampling is to integrate out trial level tensor margins $\boldsymbol{\alpha}^s_{1,t,h}, \boldsymbol{\alpha}^S_{2,t,h}, \boldsymbol{\alpha}^s_{3,t,h}$ and arrange $\mathbf{y}^s_{t,h}$ as a vector so (4.5) and (4.6) become

$$\tilde{\mathbf{y}}_{t,h} = F_{f,t,h}\boldsymbol{\alpha}_{f,t,h} + \boldsymbol{\epsilon}_{f,t,h}, \quad \boldsymbol{\epsilon}_{f,t,h} \sim \mathcal{N}(\mathbf{0}, \Sigma_{f,t,h}), \tag{4.11}$$

where

$$\tilde{\mathbf{y}}_{t,h} = \begin{bmatrix} \tilde{\mathbf{y}}^1_{t,h} \\ \vdots \\ \tilde{\mathbf{y}}^S_{t,h} \end{bmatrix}, \quad F_{f,t,h} = \begin{bmatrix} F^1_{f,t,h} \\ \vdots \\ F^S_{f,t,h} \end{bmatrix}, \quad \Sigma_{f,t,h} = \begin{bmatrix} \Sigma^1_{f,t,h} & \mathbf{0}_N & \dots & \mathbf{0}_N \\ \mathbf{0}_N & \Sigma^2_{f,t,h} & \dots & \mathbf{0}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_N & \mathbf{0}_N & \dots & \Sigma^S_{f,t,h} \end{bmatrix}.$$

and $\Sigma^s_{f,t,h} = F^s_{f,t,h}\Lambda_{f,h}{F^s_{f,t,h}}^\top + \Sigma^s$. Together with linear Gaussian state equation implied by the two priors that we investigate, it completes a conditionally Gaussian state space specification so techniques like the Kalman filter or the all without a loop (AWOL) algorithm (Rue, 2001; McCausland et al., 2011; Kastner and Frühwirth-Schnatter, 2014) can be employed to draw samples of the state variables. After updating $\boldsymbol{\alpha}^s_{1,t,h}, \boldsymbol{\alpha}^S_{2,t,h}, \boldsymbol{\alpha}^s_{3,t,h}$, new trial specific tensor margins are sampled from a multivariate normal distribution. Gibbs sampler for DSS priors in BHTVT-VAR models is straightforward while for the DSP prior, it requires extra procedures to derive the linear Gaussian state equation given certain hyperparameters. First, in order to draw posterior samples of $\mathbf{g}_{f,t,h}$ from the model, one can transform (4.9) into

$$\log\left((\boldsymbol{\alpha}_{f,t,h} - \boldsymbol{\alpha}_{f,t-1,h})^2\right) = \mathbf{g}_{f,t,h} + \log\left(\mathbf{e}^2_{f,t,h}\right), \quad \mathbf{e}_{f,t,h} \sim \mathcal{N}(0, \mathbf{I}_{r_f}),$$

where $\mathbf{I}_{r_f}$ denotes the identity matrix whose dimension is N when $f = 1, 2$ and P when $f = 3$. The log squared errors $\log(\mathbf{e}^2_{f,t,h})$ are approximated by ten-component Gaussian mixture as in Omori et al. (2007). Let $\mathbf{l}_{f,t,h}$ be a vector of length $r_f, f = 1, 2, 3$ whose entries $\mathbf{l}^i_{f,t,h} = 1, \dots, 10,\ i = 1, \dots, r_f$ are component indices of the Gaussian mixture, $q_{\mathbf{l}^i_{f,t,h}}$, $m_{\mathbf{l}^i_{f,t,h}}$ and $v_{\mathbf{l}^i_{f,t,h}}$ denote the corresponding weight, Gaussian mean and variance that are known. It is also possible to express the error term $\boldsymbol{\eta}_{f,t,h}$ in (4.10), distributed according to Z-distribution, as a mean-variance scale mixture of Gaussian. The mixing distribution is the Pólya-gamma random variable (Barndorff-Nielsen et al., 1982; Polson et al., 2013),

$$\boldsymbol{\eta}_{f,t,h} \mid \boldsymbol{\xi}_{f,t,h} \sim \mathcal{N}\left((a_\eta - b_\eta)/2 \cdot {\boldsymbol{\xi}_{f,t,h}}^{-1}, \operatorname{diag}\left\{{\boldsymbol{\xi}_{f,t,h}}^{-1}\right\}\right),$$
$$\boldsymbol{\xi}^i_{f,t,h} \sim \mathrm{PG}(a_\eta + b_\eta, 0), \quad i = 1, \dots, r_f, \quad f = 1, 2, 3.$$

Since the system is linear and the error terms can be written in terms of Gaussian mixtures, we are able to achieve fast posterior inferences in the DLM framework and through the use of the Kalman filter. Details of the Gibbs samplers are given in the appendix.

To validate the performance of the two prior specifications in simulation studies, we focus on the following criteria. Posterior inclusion probabilities $\mathrm{pr}(\boldsymbol{\gamma}_{f,t,h} = 1 \mid \mathbf{y}_{1:T}), f = 1, 2, 3, t = P + 1, \dots, T$ and $h = 1, \dots, H$ are specific to the DSS prior where binary variables $\boldsymbol{\gamma}_{f,t,h}$ indicate whether the corresponding variables are from the spike or the slab distribution. Note that $\boldsymbol{\gamma}_{f,t,h} = 1$ only reflects activation in tensor margin level, the coefficient level activation matrix $\Gamma_t \in \mathbb{R}^{N\times NP}$ can be determined by

$$\Gamma_{t,ji} = \begin{cases} 0, & \left[\sum_{h=1}^H \boldsymbol{\gamma}_{1,t,h} \circ \boldsymbol{\gamma}_{2,t,h} \circ \boldsymbol{\gamma}_{3,t,h}\right]_{ji} = 0 \\ 1, & \text{otherwise} \end{cases},$$

where $\Gamma_{t,ji}$ and $\left[\sum_{h=1}^H \boldsymbol{\gamma}_{1,t,h} \circ \boldsymbol{\gamma}_{2,t,h} \circ \boldsymbol{\gamma}_{3,t,h}\right]_{ji}$ denote the jth row ith column of the matrix Γ_t and matrix $\sum_{h=1}^H \boldsymbol{\gamma}_{1,t,h} \circ \boldsymbol{\gamma}_{2,t,h} \circ \boldsymbol{\gamma}_{3,t,h}$. For the DSP prior, $\Gamma_{t,ji}$ is defined based on posterior samples of the credible interval

$$\Gamma_{t,ji} = \begin{cases} 1, & \mathbb{1}\left(\mathrm{pr}\left(|\hat{A}_{t,ji}| > 0\right) > 0.95\right) \\ 0, & \text{otherwise} \end{cases},$$

where $\hat{A}_{t,ji}$ is the posterior estimates of the jth row and ith column of the dynamic coefficient matrix. MCMC samples of Γ_t can be used to calculate the posterior inclusion probabilities $\mathrm{pr}(\Gamma_{t,ji} = 1 \mid \mathbf{y}_{1:T})$ of each cell in the TV-VAR coefficient matrix and furthermore derivatives from a confusion table. We also compute the root mean squared errors (RMSE) for both cases of the DSS and the DSP prior

$$\mathrm{RMSE}(\hat{A}_t) = \sqrt{\frac{1}{N^2P}\sum_{t=P+1}^{T}\sum_{i=1}^{NP}\sum_{j=1}^{N}\left(\hat{A}_{t,ji} - A_{t,ji}\right)^2},$$

where $\hat{A}_t$ is the posterior estimates of the mean coefficient matrix, A_t is the true mean coefficient matrix and $A_{t,ji}$ denotes its jth row ith column entry. RMSE measures how well the posterior mean follow the true values and the uncertainty of these estimates is reflected by mean credible intervals widths (MCIW)

$$\mathrm{MCIW} = \frac{1}{N^2P}\sum_{t=P+1}^{T}\sum_{i=1}^{NP}\sum_{j=1}^{N}\left(\hat{A}_{t,ji}^{(97.5)} - \hat{A}_{t,ji}^{(2.5)}\right),$$

with $\hat{A}_{t,ji}^{(97.5)}$ and $\hat{A}_{t,ji}^{(2.5)}$ being the upper and lower bound of the 95% credible interval of the jth row ith column of the dynamic coefficient matrix estimate $\hat{A}_t$.

4.5 Simulation Studies

We consider three simulation studies. In each simulation study, data in $S = 5, 20, 80$ trials are generated from a small BHTVT-VAR model where $N = 4, P = 3$ and the PARAFAC decomposition of the stacked dynamic coefficient tensor has rank $H = 3$. The tensor margins $\boldsymbol{\alpha}_{1,t,h}, \boldsymbol{\alpha}_{2,t,h}, \boldsymbol{\alpha}_{3,t,h}$ are simulated from an AR(1) process with an autocorrelation coefficient of 0.98 as in Rockova and McAlinn (2021). After simulating tensor margins from the AR process, we set the entries of $\boldsymbol{\alpha}_{1,t,h}, \boldsymbol{\alpha}_{2,t,h}, \boldsymbol{\alpha}_{3,t,h}$ with absolute values below a 0.5 threshold to 0. This is to ensure time intervals of exact 0's in the time-varying parameters that are used to generate the time series. Trial specific coefficients $\boldsymbol{\alpha}_{1,t,h}^{s}, \boldsymbol{\alpha}_{2,t,h}^{s}, \boldsymbol{\alpha}_{3,t,h}^{s}$ are sampled from corresponding normal distributions with 0.1 standard deviation. Finally, the true coefficient matrix is computed from the tensor margins for each trial and $T = 200$ time series are generated from the TV-VAR model where the error term has a time invariant diagonal covariance matrix with diagonal entries all equal to 0.1. In this step, we check the range of simulated observations. If the range is larger than 10, new tensor margins of each trial will be generated until the condition is fulfilled. Examples of simulated time series are displayed in Figure 4.1.

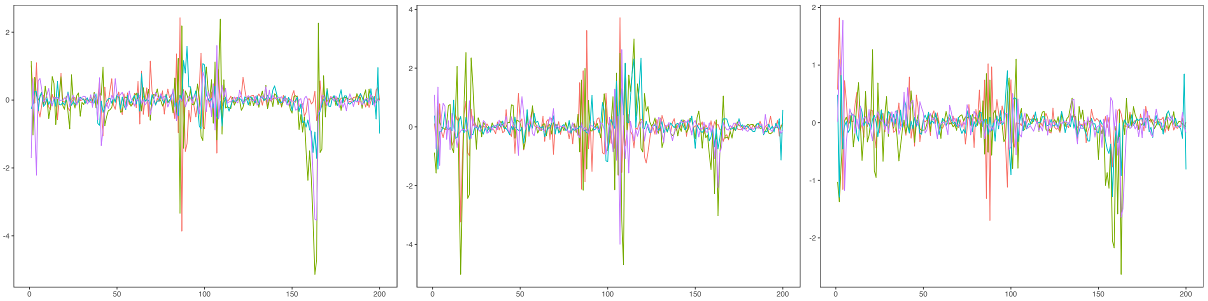


Figure 4.1: Simulated time series observations from the BHTVT-VAR model.

To conduct posterior inferences, we assume the TV-VAR model order $P = 4$. The initial value of tensor margins $\boldsymbol{\alpha}_{1,t,h}, \boldsymbol{\alpha}_{2,t,h}, \boldsymbol{\alpha}_{3,t,h}$ and trials specific tensor margins $\boldsymbol{\alpha}^s_{1,t,h}, \boldsymbol{\alpha}^s_{2,t,h}, \boldsymbol{\alpha}^s_{3,t,h}$ take the same value as the tensor margins derived from the static VAR estimates. The PARAFAC decomposition rank is chosen to be $H = 4$, larger than the true rank. In DSS prior, we set the variance of the spike normal distribution $\kappa_{f,0,h}$ and the slab normal distribution $\kappa_{f,1,h}$ to be 0.01 and 0.1. The beta prior on the autoregressive coefficients $\boldsymbol{\phi}^i_{f,1,h}$ has parameter $a_\phi = 20$ and $b_\phi = 1.5$ (Kim et al., 1998). In the DSP prior, $a_\eta = 0.5$ and $b_\eta = 0.5$ are chosen as this yields the special case of horseshoe prior. As for the same beta prior on $\boldsymbol{\phi}^i_{f,1,h}$, we follow the guideline in Kowal et al. (2019) to set $a_\phi = 10$ and $b_\phi = 2$. Lastly, σ_ϕ in the prior of $\boldsymbol{\phi}^i_{f,0,h}$ is set to 1. Parameters that are common regardless of the dynamic shrinkage prior choices, the diagonal elements of the covariance matrices Σ^s and $\Lambda_{f,h}$, have the same hyperparameter values $a_\sigma = 100, b_\sigma = 1$ and $a_\lambda = 50, b_\lambda = 1$. In all simulation studies, 10000 MCMC iterations are run. We discard the first 3000 sample and thin the remain samples by 3. Figure 4.2 compares the estimated trajectories (posterior mean) with the true trajectories of certain cells in the TV-VAR coefficient matrix. In general, both priors are able to capture changes and follow the trend of the true dynamic coefficients. In Table 4.1, we report RMSE and MCIW when imposing the DSS and DSP prior. False negative rate (FNR), false positive rate (FPR), false discovery rate (FDR) and false omission rate (FOR) are also included in case of the DSS prior. When more trials are i pooled together to make inference about the TV-VAR coefficients, the DSS prior performs better in the sense that RMSE and FNR decrease. However better performances in RMSE and FNR come at the expense of more uncertainty showing in wider credible intervals and higher FDR.

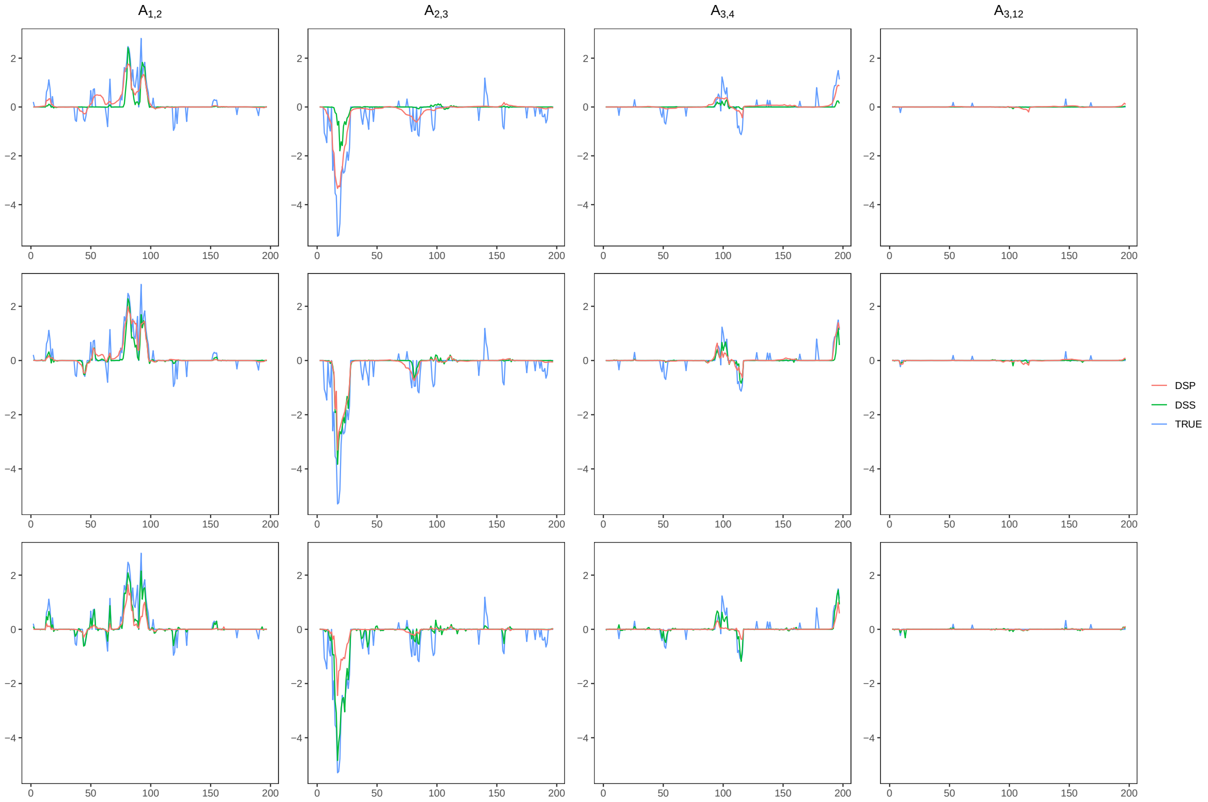


Figure 4.2: Estimated and true trajectories of selected dynamic coefficients $A_{1,2}$, $A_{2,3}$, $A_{3,4}$ and $A_{3,12}$. The estimates are posterior means of the DSP prior and the DSS prior. Top, middle and bottom rows correspond to $S = 5$, $S = 20$ and $S = 80$ respectively.

DSS	MCIW	RMSE	FNR	FPR	FDR	FOR
$\Theta = 0.1$	0.2388	0.2687	0.8841	0.0056	0.2474	0.1197
	(0.0539)	(0.0528)	(0.0458)	(0.0028)	(0.0994)	(0.0283)
$\Theta = 0.2$	0.2803	0.2518	0.7961	0.0117	0.2771	0.1096
	(0.0637)	(0.0481)	(0.0645)	(0.0047)	(0.0783)	(0.0260)
$\Theta = 0.3$	0.3203	0.2398	0.7101	0.0194	0.3055	0.0995
	(0.0680)	(0.0440)	(0.0743)	(0.0077)	(0.0746)	(0.0238)
$\Theta = 0.5$	0.5759	0.2319	0.1289	0.5372	0.8015	0.0403
	(0.0759)	(0.0407)	(0.0553)	(0.0923)	(0.0315)	(0.0177)
$\Theta = 0.7$	0.7352	0.2275	0.0000	1.000	0.8677	NaN
	(0.0840)	(0.0395)	(0.0000)	(0.0001)	(0.0287)	(NA)
$\Theta = 1$	1.0621	0.2261	0.0000	1.0000	0.8677	NaN
	(0.1294)	(0.0354)	(0.0000)	(0.0000)	(0.0287)	(NA)
DSP	MCIW	RMSE	FNR	FPR	FDR	FOR
H=3	0.7428	0.2090	0.6969	0.0057	0.1094	0.0965
$\eta_a = 0.5, \eta_b = 0.5$	(0.0651)	(0.0331)	(0.0665)	(0.0032)	(0.0563)	(0.0227)
H=4	0.9558	0.2940	0.7160	0.0042	0.0831	0.0986
$\eta_a = 0.5, \eta_b = 0.5$	(1.6116)	(0.8780)	(0.0692)	(0.0057)	(0.0858)	(0.0223)
H=4	0.4268	0.2546	0.5768	0.0352	0.3486	0.0831
$\eta_a = 0.01, \eta_b = 0.99$	(0.6495)	(0.3653)	(0.0870)	(0.0153)	(0.0885)	(0.0198)

Table 4.1: Performance evaluation of posterior estimates with BHTVT-VAR models under DSS and DSP prior specification.

4.6 Real Data Application

We apply the BHTVT-VAR model for multiple trials to LFP data recorded in the hippocampal region CA1 under two experiment conditions in a odor sequence memory task to study the role of the hippocampus in coding for the memory of sequential relationships among nonspatial events (Allen et al., 2016; Hu et al., 2020). We focus on estimating the TV-VAR coefficients under an "InSeq" condition and conducting inferences on how brain effective connectivity patterns evolve over time. In the experiment, rats were presented with repeated sequences of five odors in a single odor port. They were trained to identify whether each odor was presented "in sequence" (by holding their nose poke until the signal delivered after 1.2s) or "out of sequence" (by withdrawing their nose poke before the signal) to receive a water reward (Allen et al., 2016). The LFP data that we will be using were recorded from $N = 21$ CA1 electrodes in 245 trials/epochs, spanning roughly

4 seconds with 4000 time points. We select trials where only A, B, C, D odors were presented, and they were presented in sequence, and the rat correctly identified that each odor was in sequence. The filtering step gives us $S = 95$ trials for further analysis. In addition, we reduce the temporal resolution of time series by average over every 10 time points to $T = 400$. The PARAFAC decomposition rank is chosen to be $H = 4$ so that at least 50% of the Frobenius norm is explained by the approximation. Partial directed coherence (PDC) is a measure of brain effective connectivity that resembles the Granger causality in frequency domain. Given the condition level TV-VAR dynamic coefficient matrix $\left[A_{1,t}, A_{2,t}, \ldots, A_{P,t}\right]$, the time-varying PDC from node i to node j of a certain frequency ω at time t is defined as (Omidvarnia et al., 2013; Leistritz et al., 2013)

$$\mathrm{PDC}_{t,ji}(\omega) = \frac{|A_{t,ji}(\omega)|^2}{\sum_{j=1}^{N} |A_{t,ji}(\omega)|^2},$$

where $A_{t,ji}(\omega)$ is the jth row ith column of matrix $A_t(\omega)$ computed from the time-varying coefficients

$$A_t(\omega) = I - \sum_{p=1}^{P} A_{p,t} \exp\left(-2i\pi\omega p\right).$$

At each time point t, PDC is actually a normalized quantity since $\sum_{j=1}^{N} \mathrm{PDC}_{ji}(\omega) = 1$, therefore it measures the proportion that the outflow from i to j makes up in all the information outflow from i, but not the absolute strength of such information transmission.

We add one final remark that we are working on discovering scientific findings in the direction highlighted by Shahbaba et al. (2019) from the perspective of brain effective connectivity and preliminary results on principal component analysis (PCA) of the estimated tensor margins seem promising.

Appendix

We report Gibbs samplers used to draw from the posterior distribution of the BHTVT-VAR model under our prior specifications.

Gibbs sampler for DSS priors

1. update $\boldsymbol{\alpha}_{f,t,h}$ from the conditionally Gaussian state space model with state equation (4.7) and observation equation (4.11) using the FFBS algorithm.

2. update $\boldsymbol{\alpha}^s_{f,t,h}$ from the multivariate normal distribution

$$\mathcal{N}\left({\Sigma^s_{f,t,h}}^*\left({\Lambda_{f,h}}^{-1}\boldsymbol{\alpha}_{f,t,h}+{F^s_{f,t,h}}^\top(\Sigma^s)^{-1}\mathbf{y}^s_{t,h}\right),\, {\Sigma^s_{f,t,h}}^*\right),$$

   where ${\Sigma^s_{f,t,h}}^* = \left({F^s_{f,t,h}}^\top(\Sigma^s)^{-1}F^s_{f,t,h}+{\Lambda_{f,h}}^{-1}\right)^{-1}$.

3. update $\boldsymbol{\gamma}_{f,t,h}$ by first computing $\boldsymbol{\theta}_{f,t,h}$ according to (4.8), then computing

$$\mathbf{p}_{f,t,h} = \frac{\boldsymbol{\gamma}_{f,t,h}\circ\psi_1(\boldsymbol{\alpha}^s_{f,t,h}\mid\boldsymbol{\mu}_{f,t,h},\,\kappa_{f,1,h})}{\boldsymbol{\gamma}_{f,t,h}\circ\psi_1(\boldsymbol{\alpha}^s_{f,t,h}\mid\boldsymbol{\mu}_{f,t,h},\,\kappa_{f,1,h})+(1-\boldsymbol{\gamma}_{f,t,h})\circ\psi_0(\boldsymbol{\alpha}^s_{f,t,h}\mid\kappa_{f,0,h})}$$

   as parameters of the Bernoulli distribution to sample $\boldsymbol{\gamma}_{f,t,h}$.

4. update $\boldsymbol{\phi}_{f,1,h}$ with a Metropolis Hastings step.

5. update diagonal entries of $\Lambda_{f,h}$ from

$$\text{IG}\left(a_\lambda + S(T-P)/2,\, b_\lambda + \sum_{t=P+1}^{T}\sum_{s=1}^{S}\|\boldsymbol{\alpha}^s_{f,t,h}-\boldsymbol{\alpha}_{f,t,h}\|_2^2\right).$$

6. update diagonal entries of Σ^s from

$$\mathrm{IG}\left(a_\sigma + (T-P)/2,\, b_\sigma + \left\| \mathbf{y}_t^s - \sum_{h=1}^{H} {\boldsymbol{\alpha}_{3,t,h}^s}^\top \otimes \left(\boldsymbol{\alpha}_{1,t,h}^s \circ \boldsymbol{\alpha}_{2,t,h}^s\right) \begin{bmatrix} \mathbf{y}_{t-1}^s \\ \vdots \\ \mathbf{y}_{t-P}^s \end{bmatrix} \right\|_2^2 \right).$$

Gibbs sampler for DSP priors

1. update $\boldsymbol{\alpha}_{f,t,h}$ from the conditionally Gaussian state space model with observation equation (4.11) and state equation

$$\boldsymbol{\alpha}_{f,t,h} = G^{DSP}\boldsymbol{\alpha}_{f,t,h} + \boldsymbol{\omega}_{f,t,h}, \quad \boldsymbol{\omega}_{f,t,h} \sim \mathcal{N}(\mathbf{0}, W_{f,t,h}^{DSP})$$

with G^{DSP} an identity matrix of appropriate dimension and $W_{f,t,h}^{DSP} = \mathrm{diag}\{e^{\mathbf{g}_{f,t,h}}\}$ using the FFBS algorithm.

2. update $\boldsymbol{\alpha}_{f,t,h}^s$ from the multivariate normal distribution

$$\mathcal{N}\left({\Sigma_{f,t,h}^s}^* \left({\Lambda_{f,h}}^{-1}\boldsymbol{\alpha}_{f,t,h} + {F_{f,t,h}^s}^\top (\Sigma^s)^{-1}\mathbf{y}_{t,h}^s\right),\, {\Sigma_{f,t,h}^s}^*\right),$$

where ${\Sigma_{f,t,h}^s}^* = \left({F_{f,t,h}^s}^\top (\Sigma^s)^{-1} F_{f,t,h}^s + {\Lambda_{f,h}}^{-1}\right)^{-1}$.

3. update $\mathbf{g}_{f,t,h}$ from the DLM with observation equation conditional on component index $\mathbf{l}_{f,t,h}$

$$\log\left((\boldsymbol{\alpha}_{f,t,h} - \boldsymbol{\alpha}_{f,t-1,h})^2\right) = \mathbf{g}_{f,t,h} + {\mathbf{e}_{f,t,h}}^*$$

$${\mathbf{e}_{f,t,h}}^* \sim \mathcal{N}\left(\begin{bmatrix} m_{\mathbf{l}_{f,t,h}^1} \\ \vdots \\ m_{\mathbf{l}_{f,t,h}^{r_f}} \end{bmatrix}, \begin{bmatrix} v_{\mathbf{l}_{f,t,h}^1} & 0 & \dots & 0 \\ 0 & v_{\mathbf{l}_{f,t,h}^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & v_{\mathbf{l}_{f,t,h}^{r_f}} \end{bmatrix}\right),$$

and state equation

$$\mathbf{g}_{f,t,h} = \boldsymbol{\phi}_{f,0,h} + \boldsymbol{\phi}_{f,1,h} \circ (\mathbf{g}_{f,t-1,h} - \boldsymbol{\phi}_{f,0,h}) + \boldsymbol{\eta}_{f,t,h},$$

where $\boldsymbol{\eta}_{f,t,h} \sim \mathcal{N}\left((a_\eta - b_\eta)/2 \cdot {\boldsymbol{\xi}_{f,t,h}}^{-1},\, \text{diag}\left\{{\boldsymbol{\xi}_{f,t,h}}^{-1}\right\}\right)$ using the AWOL algorithm.

4. update $\mathbf{l}_{f,t,h}$ from the multinomial distribution with probabilities of $l = 1, \ldots, 10$ determined by

$$\text{pr}(\mathbf{l}_{f,t,h} = l) \frac{\mathcal{N}\left(\log\left((\boldsymbol{\alpha}_{f,t,h} - \boldsymbol{\alpha}_{f,t-1,h})^2\right) - \mathbf{g}_{f,t,h} \mid m_l,\, v_l\right)}{\sum_{l=1}^{10} \mathcal{N}\left(\log\left((\boldsymbol{\alpha}_{f,t,h} - \boldsymbol{\alpha}_{f,t-1,h})^2\right) - \mathbf{g}_{f,t,h} \mid m_l,\, v_l\right)},$$

   here we abuse the notation $\mathcal{N}$ to mean the Gaussian density value.

5. update $\boldsymbol{\xi}_{f,t,h}$ from $\text{PG}\left(a_\eta + b_\eta,\, \mathbf{g}_{f,t,h} - \boldsymbol{\phi}_{f,0,h} - \boldsymbol{\phi}_{f,1,h} \circ (\mathbf{g}_{f,t-1,h} - \boldsymbol{\phi}_{f,0,h})\right)$.

6. update $\boldsymbol{\phi}_{f,1,h}$ with a slice sampler (Neal, 2003) or a Metropolis Hastings step.

7. update $\boldsymbol{\phi}_{f,0,h}$ from $\mathcal{N}(Q_{f,h}\mathbf{c}_{f,h}, Q_{f,h})$, where

$$Q_{f,h} = \text{diag}\left\{\boldsymbol{\xi}_{f,P,h} + \boldsymbol{\xi}_{\phi,f,h} + (1 - \boldsymbol{\phi}_{f,1,h})^2 \circ \sum_{t=P+1}^{T} \boldsymbol{\xi}_{f,t,h}\right\}^{-1}$$

   and

$$\mathbf{c}_{f,h} = \log\left(\frac{\sigma_\phi^2}{T-P}\right)\boldsymbol{\xi}_{\phi,f,h} + \boldsymbol{\xi}_{f,P,h} \circ \mathbf{g}_{f,1,h} + (1 - \boldsymbol{\phi}_{f,1,h}) \circ \sum_{t=P+1}^{T} \boldsymbol{\xi}_{f,t,h} \circ (\mathbf{g}_{f,t,h} - \boldsymbol{\phi}_{f,1,h} \circ \mathbf{g}_{f,t-1,h}).$$

8. update $\boldsymbol{\xi}_{\phi,f,h}$ from $\text{PG}\left(1,\, \boldsymbol{\phi}_{f,0,h} - \log\left(\sigma_\phi^2 / (T-P)\right)\right)$.

9. update diagonal entries of $\Lambda_{f,h}$ from

$$\text{IG}\left(a_\lambda + S(T-P)/2,\, b_\lambda + \sum_{t=P+1}^{T} \sum_{s=1}^{S} \|\boldsymbol{\alpha}_{f,t,h}^{s} - \boldsymbol{\alpha}_{f,t,h}\|_2^2\right).$$

10. update diagonal entries of Σ^s from

$$\mathrm{IG}\left(a_\sigma + (T-P)/2, b_\sigma + \left\| \mathbf{y}_t^s - \sum_{h=1}^{H} {\boldsymbol{\alpha}_{3,t,h}^s}^\top \otimes \left(\boldsymbol{\alpha}_{1,t,h}^s \circ \boldsymbol{\alpha}_{2,t,h}^s\right) \begin{bmatrix} \mathbf{y}_{t-1}^s \\ \vdots \\ \mathbf{y}_{t-P}^s \end{bmatrix} \right\|_2^2 \right).$$

Chapter 5

Dynamic Generalized Double Pareto Process Priors

In this chapter, we develop and investigate an alternative to the dynamic variable selection priors discussed in Chapter 4, which is based on the use of a Generalized Double Pareto (GDP) process. We motivate the proposed process prior in a simple dynamic regression framework,

$$y_t = \beta_t x_t + \epsilon_t, \quad \epsilon \sim \mathcal{N}(0, \sigma_t^2), \tag{5.1}$$

y_t and $x_t, t = 1, \ldots, T$ are the observed response variable and covariate. Extensions to the tensor TVT-VAR models can be made following the same ideas presented in the Chapter 4. Here, β_t denotes the dynamic regression coefficient and we assume that if follows a GDP distribution, characterized by the following hierarchical representation,

$$\begin{aligned}
\beta_t \mid \tau_t &\sim \mathcal{N}(0, \tau_t), \\
\tau_t \mid \lambda_t &\sim \text{Exp}(\lambda_t^2/2), \\
\lambda_t &\sim \text{Ga}(\alpha, \eta),
\end{aligned}$$

i.e., as a scale mixture of Normals, similarly to the DSS and DSP priors of Chapter 4. We illustrate the GDP process prior using this specification. The idea is to keep the GDP prior as the marginal distribution of β_t for any t, but in the meanwhile to allow for a clear dependence structure, especially for τ_t which is the crucial parameter that shrinks β_t towards 0. For instance, if the shrinkage is strong at time t, it may follow naturally to also expect a strong shrinkage at time $t+1$. We start by noting that rhe probability density function of τ_t, after integrating out λ_t, is

$$\begin{aligned}\mathrm{pr}(\tau_t \mid \alpha, \eta) = \frac{2^{\alpha/2-1}\eta^{\alpha}}{\Gamma(\alpha)\tau_t^{(\alpha+3)/2}} \Bigg(\tau_t^{1/2}\Gamma\left(\frac{\alpha+2}{2}\right) {}_1F_1\left(\frac{\alpha+2}{2}, \frac{1}{2}, \frac{\eta^2}{2\tau_t}\right) - \\ \sqrt{2}\eta\Gamma\left(\frac{\alpha+3}{2}\right) {}_1F_1\left(\frac{\alpha+3}{2}, \frac{3}{2}, \frac{\eta^2}{2\tau_t}\right)\Bigg), \qquad (5.2)\end{aligned}$$

where ${}_1F_1(\cdot)$ is the confluent hypergeometric function. With this density function, it is difficult to construct a dependence structure on τ_t while maintaining a stationary time-series. To circumvent this challenge, we propose modeling the dependence structure on τ_t conditioning on a fixed λ_t, so that $\mathrm{pr}(\tau_t \mid \lambda)$ simplifies to $\mathrm{Exp}(\lambda^2/2)$. Exponential AR(1) processes are suitable choices to introduce temporal dependence on τ_t given λ, by assuming

$$\tau_t = \rho\tau_{t-1} + \omega_t, \qquad (5.3)$$

for some $0 < \rho < 1$. Under the assumption of stationarity, this can be seen as a particular case of the general problem of self-decomposability. A random variable X is self-decomposable if for any $0 < \rho < 1$, there is an independent random variable $X(\rho)$ such that

$$X = \rho X + X(\rho).$$

In general, the characteristic function of $X(\rho)$ is equal to the ratio between the characteristic function of X and ρX since X and $X(\rho)$ are independent. Finding such an error term $X(\rho)$ is challenging for many distributions. However, Gaver and Lewis (1980) show that the error term ω_t in (5.3) is exactly 0 with probability ρ; otherwise, with probability $1-\rho$, the term follows the exponential distribution with mean $2/\lambda^2$.

A possible modeling choice is the Normal-Gamma autoregressive (NGAR) by Kalli and Griffin (2014), which has then been used by Rockova and McAlinn (2021) to define the Laplace AR process. The resulting dependence structure is a first order autoregressive process in the sense that $E(\tau_t \mid \tau_{t-1})$ is linear in τ_{t-1} with autoregressive coefficient ρ. The AR dependence is accomplished by introducing an auxiliary random variables ψ_t, defined as

$$\begin{aligned} \psi_t \mid \tau_{t-1} &\sim \text{Ga}(1+\tau_{t-1}, \lambda^2/(2(1-\rho))) \\ \tau_{t-1} \mid \psi_t &\sim \text{Pois}(\rho\lambda^2\psi_{t-1}/2(1-\rho)). \end{aligned}$$

such that the marginal distribution of τ_t is $\text{Exp}(\lambda^2/2)$ (Pitt et al., 2002; Pitt and Walker, 2005).

In this chapter, we investigate a third option which is motivated by the fact that the time varying τ_t should reflect stochastic volatility, and compare this approach with those above. A common approach to deal with volatility in time-series is to transform the original problem into a linear problem by considering

$$\log(\beta_t^2) = \theta_t + \log(e^2) \tag{5.4}$$

where $\theta_t = \log(\tau_t)$ and $e \sim \mathcal{N}(0,1)$. Hence, instead of assuming τ_t as an AR(1) process, one can introduce temporal dependence by requiring θ_t to be a stationary AR(1) while preserving the marginal distribution in (5.2). Thus, it is necessary to identify the distribution of ω_t such that

$$\theta_t = \rho\,\theta_{t-1} + \omega_t,$$

admits invariant marginal distribution for τ_t as above, with $0 < \rho < 1$ since the shrinkage should be positively correlated. Conditional on λ, $\log(\tau_t)$ is a shifted Gumbel distribution.

We conclude this Section by noticing that if one further desires to model the dependence between β_{t-1} and β_t, it can be assumed that jointly β_{t-1} and β_t follow a bivariate normal distribution such that

$$\begin{pmatrix} \beta_{t-1} \\ \beta_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{t-1} & \phi\sqrt{\tau_{t-1}\tau_t} \\ \phi\sqrt{\tau_{t-1}\tau_t} & \tau_t \end{pmatrix} \right),$$

which implies that $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\phi\sqrt{\tau_t/\tau_{t-1}}\beta_{t-1}, (1-\rho^2)\tau_t)$. The previous result can be written in linear form as

$$\beta_t = \phi\sqrt{\tau_t/\tau_{t-1}}\beta_{t-1} + \nu_t \quad \nu_t \sim \mathcal{N}(0, (1-\phi^2)\tau_t).$$

The process of β_t combined with the first order AR(1) process of τ_t is a special case of the NGAR proposed by Kalli and Griffin (2014). However, by proceeding this way, it is not possible to obtain an expression similar to (5.4). Later, we argue that the different ways we have discussed to construct the dependence on τ_t indeed correspond to incorporating different prior information in the modeling of the evolution of the sparsity of the regression coefficients.

5.1 A Gumbel AR(1) process

In this Section, we describe in detail the Gumbel AR(1) process of θ_t. In the following, when we write $\tau_t \sim \exp(\lambda^2/2)$ and $\theta_t = \rho\theta_{t-1} + \omega_t$, we implicitly assume that we are conditioning on λ. We start by noting that when $\tau_t \sim \exp(\lambda^2/2)$, then $-\log(\lambda^2/2) - \log(\tau_t)$ follows the standard Gumbel distribution. In this representation, the error term ω_t is related to the stable distribution S which is defined through its characteristic function

$$\psi_S(s) = \begin{cases} \exp(is\mu - |cs|^\alpha(1 - i\beta \mathrm{sgn}(s) \tan \frac{\pi\alpha}{2})) & \text{if } \alpha \neq 1 \\ \exp(is\mu - |cs|^\alpha(1 + i\beta \mathrm{sgn}(s) \frac{2}{\pi} \log |cs|)) & \text{if } \alpha = 1 \end{cases} \tag{5.5}$$

We denote the stable distribution by $\mathcal{S}(\alpha, \beta, c, \mu)$. The parameter $0 < \alpha \leq 2$ is usually called the characteristic exponent which defines the tail behavior while the parameter $-1 \leq \beta \leq 1$ governs skewness. The parameters $c > 0$ and $\mu \in \mathbb{R}$ are the scale and location parameters of the distribution, respectively. The density of the stable distribution generally does not have analytical forms unless in some special cases. For instance, when $\alpha = 2$, it simplifies to the normal distribution and the Cauchy when $\alpha = 1, \beta = 0$. Another special case is for $\beta = 1$ and $0 < \alpha < 1$. In this case, the support of the distribution is the positive real line and we obtain the so-called positive stable distributions, whose Laplace transform is $E(e^{-uS}) = e^{-u^\alpha c^\alpha / \cos(\alpha\pi/2)}$. An analytical expression of the density exists only under $\alpha = 1/2$ and the resulting distribution is the Lévy distribution, a special case of inverse Gamma distribution. The following Proposition 2 shows that to construct Gumbel distributed AR(1) processes, the error term is indeed related to a positive stable random variable.

Proposition 2. *If a positive α-stable random variable S_t is defined through its Laplace transform $E(e^{-uS_t}) = e^{-u^\alpha}, u > 0$, the AR(1) process with $0 < \alpha < 1$*

$$X_t = \alpha X_{t-1} + \alpha \log(S_t)$$

has stationary distribution $X_t \sim Gumbel(0, 1)$.

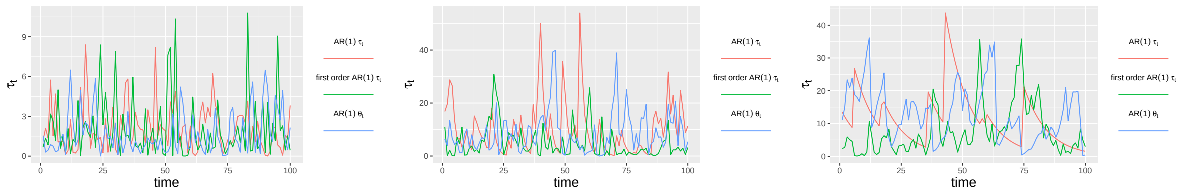
Proof. The proof follows from results in Shanbhag and Sreehari (1977); Hougaard (1986), who found that if Z and S are independent random variables such that $Z \sim \text{Exp}(1)$ and S has the Laplace transform $E(e^{-uS}) = e^{-u^\alpha}$, then the ratio $Z\,S^{-1} \sim \text{Weibull}(1, \alpha)$ coinciding with the distribution of $Z^{1/\alpha}$. Consequently we have $E(S^q) = E(Z^{-q/\alpha})/E(Z^{-q}) = \Gamma(1-q/\alpha)/\Gamma(1-q)$ and the characteristic function $\psi_{\log(S)}(s) = E(e^{is\log(S)}) = E(S^{is}) = \Gamma(1-is/\alpha)/\Gamma(1-is)$. It is known that the characteristic function of the standard Gumbel distribution is $\Gamma(1-is)$, therefore $\alpha X_{t-1} + \alpha\log(S_t)$ also has such a characteristic function due to the independence between X_{t-1} and S_t. □

The above positive stable random variable corresponds to $0 < \alpha < 1, \beta = 1, c = \cos(\alpha\pi/2)^{1/\alpha}$ and $\mu = 0$ with characteristic function $\psi_S(s) = e^{-\exp(-i\cdot\mathrm{sgn}(s)\alpha\pi/2)|s|^\alpha}$. It is easy to see that $\omega_t = -\rho\log(S_t) - (1-\rho)\log(\lambda^2/2)$ in order for θ_t to fulfill the autoregressive property.

Conditional on λ, the correlation between θ_t and θ_{t-1} is ρ due to the straightforward AR(1) construction. However, one naturally wonders what correlation between τ_t and τ_{t-1} is induced since our interest is on τ_t as it directly reflects the shrinkage effect. The relationship between τ_t and τ_{t-1} implied by AR(1) on θ_t is

$$\tau_t = e^{\omega_t}\tau_{t-1}^{\rho},$$

reflecting a non-linear dependence with dependence strength controlled by ρ. One reason to favor the modeling of θ_t instead of τ_t is that if τ_t is assumed to be an exponential AR(1) process, τ_t will be equal to $\rho\tau_{t-1}$ with positive probability by construction. Indeed, as outlined in the Introduction to this Chapter, it follows from the results by Gaver and Lewis (1980) that if $\tau_t \mid \lambda \sim \mathrm{Exp}(\lambda^2/2)$, the error term of exponential AR(1) for τ_t is exactly 0 with probability ρ and otherwise with probability $1-\rho$ it is an exponential distribution with mean $2/\lambda^2$. Figure 5.1 shows trajectories of τ_t modeled as an exponential AR(1) (red lines), as a first order AR(1) (green lines) and as the induced transformation from a Gumbel AR(1) process on θ_t (the proposed model, blue lines). It can be clearly seen that, when τ_t follows an exponential AR(1) process, the probability the the error term is equal to zero increases, leading to long segments of exponential (deterministic) decaying



(a) $\rho = 0.1$ (b) $\rho = 0.5$ (c) $\rho = 0.9$

Figure 5.1: Trajectories of τ_t under (a) small correlation, (b) medium correlation and (c) large correlation derived from exponential AR(1) processes (red lines), first order exponential AR(1) processes (green lines) and Gumbel AR(1) processes (blue lines). λ is randomly generated from Ga$(2, 1)$ and it is equal to 0.9632 in (a), 0.5246 in (b) and 0.3860 in (c).

(red) trajectories in Figure 5.1(c). Arguably, this behavior is not desirable for modeling the sparsity of dynamic regression coefficients.

5.2 Dependence between shrinkage profiles

Suppose that $x_t = 1$ and $\sigma_t^2 = 1$ in (5.1), with the prior $\beta_t \mid \tau_t \sim \mathcal{N}(0, \tau_t)$, the posterior distribution of β_t is still a normal distribution with mean $(1 - 1/(1 + \tau_t))\beta_t$ and variance $1 - 1/(1 + \tau_t)$. The quantity $\kappa_t = 1/(1 + \tau_t)$ is the shrinkage profile. It is important as it controls the level of shrinkage of the observation β_t towards zero Carvalho et al. (2010). When κ_t is around zero, the posterior mean will be be close to β_t, inducing minimal shrinkage. On the contrary, $\kappa_t \approx 1$ suggests a strong shrinkage. Since the GDP process prior preserves the marginal distribution of τ_t, the merits of the GDP prior discussed in Armagan et al. (2013) are all retained. We focus on the dependence between κ_{t-1} and κ_t. In Figure 5.2, we display the histogram of κ_{t-1} and κ_t drawn from the GDP process prior. Although the correlation ρ appears in the AR process of θ_t, its effect persists to κ_t thought τ_t so that we still have strong correlation between κ_{t-1} and κ_t. Notice that samples are not equally distributed on two sides of the diagonal line $\kappa_t = \kappa_{t-1}$. Masses are more concentrated on part of the plane above the line, meaning that κ_t is more likely to be larger than κ_{t-1}. We argue that this is an advantageous feature. When β_t stands for signal, corresponding to $\kappa_{t-1} \approx 0$, β_{t+1} can behave oppositely and become noise. On the other hand, β_t being noise ($\kappa_t \approx 1$) implies that β_{t+1} is more like to be noise as well. The GDP process prior enables shrinkage to move through time while discourages persistent signals.

5.3 Posterior Inferences

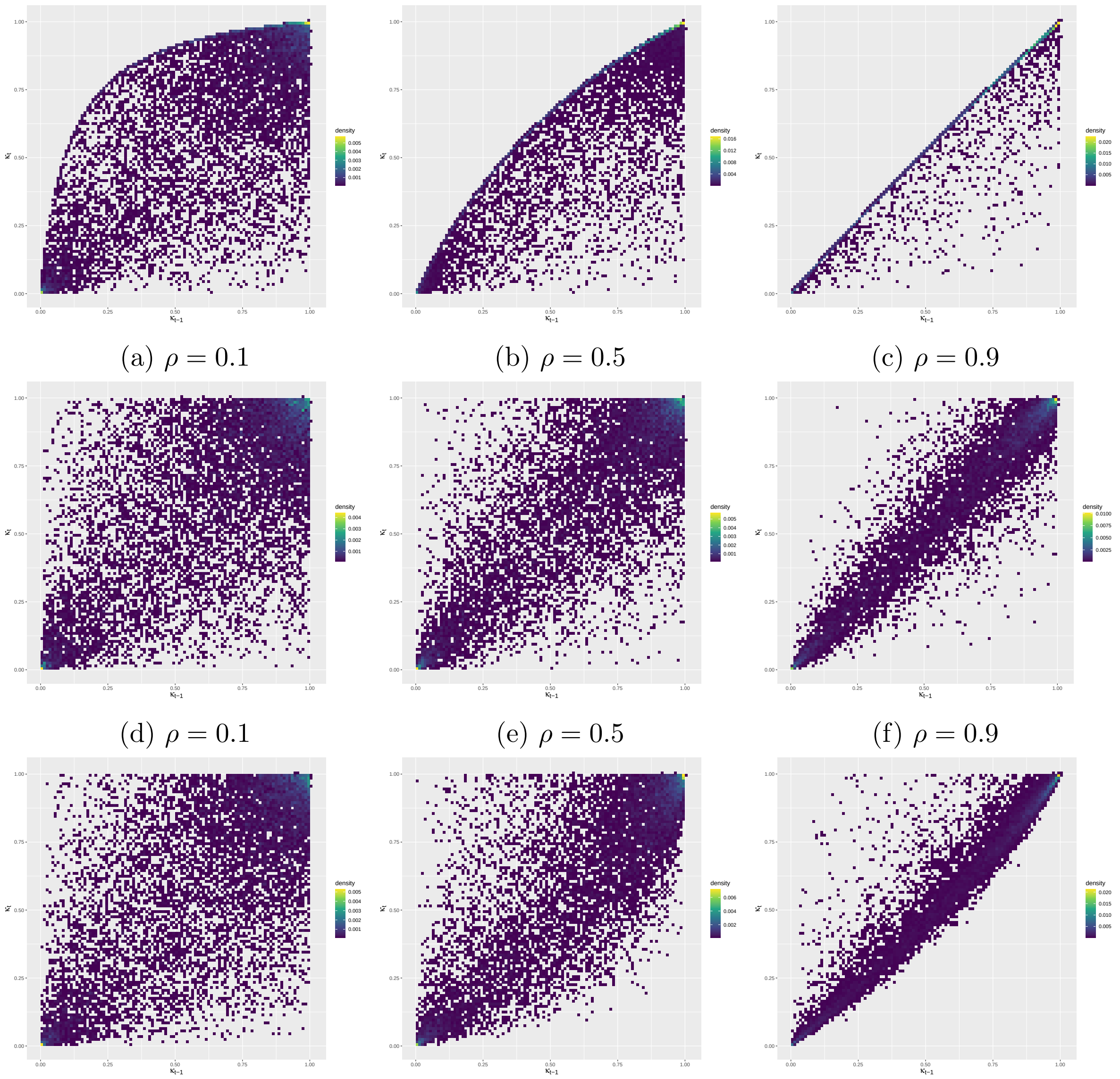
We apply a Metropolis-Hastings within Gibbs algorithm to draw posterior samples. The algorithm below is designed for the situation where β_{t-1} and β_t are correlated. First to sample the dynamic coefficients β_t given the observation and the state equations

$$\begin{aligned} y_t &= x_t\beta_t + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2), \\ \beta_t &= \phi\sqrt{\tau_t/\tau_{t-1}}\beta_{t-1} + \nu_t \quad \nu_t \sim \mathcal{N}(0, (1-\phi^2)\tau_t). \end{aligned}$$

with the initial condition that $\beta_0 \sim \mathcal{N}(0, \tau_0)$, one can use either the Forward Filtering Backward Sampling (FFBS) algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994; Shephard, 1994) or the all without a loop (AWOL) method (Rue, 2001; McCausland et al., 2011; Kastner and Frühwirth-Schnatter, 2014). The next block in the sampler is to draw θ_t from

$$\begin{aligned} \beta_t &= \phi\sqrt{\frac{e^{\theta_t}}{e^{\theta_t - 1}}}\beta_{t-1} + \nu_t \quad \nu_t \sim \mathcal{N}(0, (1-\phi^2)e^{\theta_t}), \\ \theta_t &= \rho\theta_{t-1} + \omega_t \quad \omega_t = -\rho\log(S_t) - (1-\rho)\log(\lambda^2/2). \end{aligned}$$

Adding the initial condition $\theta_0 \sim -\text{Gumbel}(0,1) - \log(\lambda^2/2)$ completes the state space representation. This is a non-Gaussian non-linear state space model where only the observation equation ν_t is Gaussian. Since FFBS or AWOL algorithms are not directly applicable to sampling $\theta_{0:T}$ from the posterior $\text{pr}(\theta_{0:T} \mid \beta_{1:T}^2, \phi, \rho, \lambda)$, we iteratively apply



(a) $\rho = 0.1$

(b) $\rho = 0.5$

(c) $\rho = 0.9$

(d) $\rho = 0.1$

(e) $\rho = 0.5$

(f) $\rho = 0.9$

(g) $\rho = 0.1$

(h) $\rho = 0.5$

(i) $\rho = 0.9$

Figure 5.2: 2-dimension histogram of the shrinkage profile κ_{t-1} and κ_t under (a) (d) (g) small correlation, (b) (e) (h) medium correlation and (c) (f) (i) large correlation. (a) (b) (c) refer to exponential AR(1) processes, (d) (e) (f) refer to first order exponential AR(1) processes and (g) (h) (i) are transformed from Gumbel AR(1) processes. $\alpha = 2$ and $\eta = 1$.

the Metropolis-Hastings algorithm as in Geweke and Tanizaki (2001) to draw samples from $\mathrm{pr}(\theta_t \mid \theta_{0:t-1}, \theta_{t+1:T}, \beta_{0:T}, \phi, \rho, \lambda)$. To make the algorithm work, evaluating $\mathrm{pr}(\theta_t \mid \theta_{t-1})$ is needed, however it is difficult as analytical expression of the density of a positive stable random variable is not available. The decomposition in Ibragimov and Chernin (1959); Kanter (1975); Simon (2011) allows us to construct a location mixture representation of $\log(S_t)$ with explicit mixture density

$$\log(S_t) = \rho^{-1}\log(b_\rho(U_t)) + (\rho - 1)/\rho \log(L_t),$$

where $b_\rho(u) = (\sin(\rho u)/\sin u)^\rho (\sin((1-\rho)u)/\sin u)^{1-\rho}$, U_t is uniform on interval $(0, \pi)$ and L_t is standard exponential distribution independent of U_t. The convolution plus the location-scale natural of the Gumbel distribution gives us the following hierarchical representation

$$\begin{aligned} \log(S_t) &\sim \mathrm{Gumbel}(\rho^{-1}\log(b_\rho(U_t)), (1-\rho)/\rho) \\ U_t &\sim \mathrm{Unif}(0, \pi). \end{aligned}$$

Given U_t, the first layer of the hierarchical structure is Gumbel distribution, whose density function is known. Through the augmented random variable U_t, $\mathrm{pr}(\theta_t \mid \theta_{t-1})$ becomes easy to evaluate.

The observation error variance σ_t^2 can be modeled either time dependent or time invariant, but it does not add conceptual difficulty to the posterior inference. In the case of homoscedasticity, one can assume the conjugate inverse Gaussian prior on σ^2. Otherwise, conventional stochastic volatility framework can be applied to account for the temporal evolution of σ_t^2. Lastly, the hyperparameter λ, autoregressive coefficients ϕ, ρ and the auxiliary variable U_t are updated via Metropolis-Hastings algorithm as well.

References

Allen, T. A., Salz, D. M., McKenzie, S., and Fortin, N. J. (2016). Nonspatial sequence coding in ca1 neurons. *Journal of Neuroscience*, 36(5):1547–1563.

Ansley, C. F. and Kohn, R. (1986). A note on reparameterizing a vector autoregressive moving average model to enforce stationarity. *Journal of Statistical Computation and Simulation*, 24(2):99–106.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119.

Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156.

Baker, A. P., Brookes, M. J., Rezek, I. A., Smith, S. M., Behrens, T., Probert Smith, P. J., and Woolrich, M. (2014). Fast transient networks in spontaneous human brain activity. *eLife*, 3:e01867–e01867.

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 145–159.

Bassett, D. S. and Bullmore, E. (2006). Small-world brain networks. *The neuroscientist*, 12(6):512–523.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

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

Billio, M., Casarin, R., Kaufmann, S., and Iacopini, M. (2018). Bayesian dynamic tensor regression. *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No*, 13.

Birn, R. M., Murphy, K., and Bandettini, P. A. (2008). The effect of respiration variations on independent component analysis results of resting state functional connectivity. *Human brain mapping*, 29(7):740–750.

Borsook, D., Maleki, N., and Burstein, R. (2015). Chapter 42 - migraine. In Zigmond, M. J., Rowland, L. P., and Coyle, J. T., editors, *Neurobiology of Brain Disorders*, pages 693–708. Academic Press, San Diego.

Canova, F. (2011). *Methods for applied macroeconomic research.* Princeton university press.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Casas, I. and Fernandez-Casal, R. (2019). tvreg: Time-varying coefficients linear regression for single and multi-equations in r. Technical report, SSRN. R package version 0.5.4.

Casas, I. and Fernandez-Casal, R. (2021). *tvReg: Time-Varying Coefficients Linear Regression for Single and Multi-Equations*. R package version 0.5.4.

Casas, I., Ferreira, E., and Orbe, S. (2017). Time-varying coefficient estimation in sure models. application to portfolio management. *Journal of Financial Econometrics*.

Chan, J. C., Eisenstat, E., and Strachan, R. W. (2020). Reducing the state space dimension in a large tvp-var. *Journal of Econometrics*, 218(1):105–118.

Chiang, S., Guindani, M., Yeh, H. J., Haneef, Z., Stern, J. M., and Vannucci, M. (2017). Bayesian vector autoregressive model for multi-subject effective connectivity inference using multi-modal neuroimaging data. *Human brain mapping*, 38(3):1311–1332.

Cichocki, A., Lee, N., Oseledets, I., Phan, A.-H., Zhao, Q., and Mandic, D. P. (2016). Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations and Trends® in Machine Learning*, 9(4-5):249–429.

Cogley, T., Primiceri, G. E., and Sargent, T. J. (2010). Inflation-gap persistence in the us. *American Economic Journal: Macroeconomics*, 2(1):43–69.

Cogley, T. and Sargent, T. J. (2005). Drifts and volatilities: monetary policies and outcomes in the post wwii us. *Review of Economic dynamics*, 8(2):262–302.

Corbin, N., Todd, N., Friston, K. J., and Callaghan, M. F. (2018). Accurate modeling of temporal correlations in rapidly sampled fmri time series. *Human brain mapping*, 39(10):3884–3897.

Dai, B., Ding, S., Wahba, G., et al. (2013). Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483.

Daunizeau, J., David, O., and Stephan, K. E. (2011). Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage*, 58(2):312–322.

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.

Duffau, H. (2012). Chapter 6 - cortical and subcortical brain mapping. In Quiñones-Hinojosa, A., editor, *Schmidek and Sweet Operative Neurosurgical Techniques (Sixth Edition)*, pages 80–93. W.B. Saunders, Philadelphia, sixth edition edition.

Durante, D. (2017). A note on the multiplicative gamma process. *Statistics & Probability Letters*, 122:198–204.

Durante, D. and Guindani, M. (2020). *Bayesian Methods in Brain Networks*, pages 1–10. American Cancer Society.

Figueroa-Jiménez, M. D., Cañete-Massé, C., Carbó-Carreté, M., Zarabozo-Hurtado, D., and Guàrdia-Olmos, J. (2021). Structural equation models to estimate dynamic effective connectivity networks in resting fmri. a comparison between individuals with down syndrome and controls. *Behavioural brain research*, 405:113188.

Flandin, G. and Penny, W. D. (2007). Bayesian fmri data analysis with sparse spatial basis function priors. *NeuroImage*, 34(3):1108–1125.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain Connectivity*, 1(1):13–36.

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302.

Friston, K. J., Preller, K. H., Mathys, C., Cagnan, H., Heinzle, J., Razi, A., and Zeidman, P. (2019). Dynamic causal modelling revisited. *Neuroimage*, 199:730–744.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2):183–202.

Gaver, D. P. and Lewis, P. (1980). First-order autoregressive gamma sequences and point processes. *Advances in Applied Probability*, pages 727–745.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Geweke, J. and Tanizaki, H. (2001). Bayesian estimation of state-space models using the metropolis–hastings algorithm within gibbs sampling. *Computational statistics & data analysis*, 37(2):151–170.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2):436–451.

Gohel, S. R. and Biswal, B. B. (2015). Functional integration between brain regions at rest occurs in multiple-frequency bands. *Brain connectivity*, 5(1):23–34.

Gorrostieta, C., Fiecas, M., Ombao, H., Burke, E., and Cramer, S. (2013). Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity. *Frontiers in Computational Neuroscience*, 7:159.

Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Granger, C. W. (2008). Non-linear models: Where do we go next-time varying parameter models? *Studies in Nonlinear Dynamics & Econometrics*, 12(3).

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1):2733–2763.

Haslbeck, J. M., Bringmann, L. F., and Waldorp, L. J. (2021). A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behavioral Research*, 56(1):120–149.

Haufe, S., Nikulin, V. V., Müller, K.-R., and Nolte, G. (2013). A critical assessment of connectivity measures for eeg data: a simulation study. *Neuroimage*, 64:120–133.

Herlin, B., Navarro, V., and Dupont, S. (2021). The temporal pole: from anatomy to function-a literature appraisal. *Journal of Chemical Neuroanatomy*, page 101925.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169.

Hougaard, P. (1986). A class of multivanate failure time distributions. *Biometrika*, 73(3):671–678.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications.* Sage.

Hsu, N.-J., Hung, H.-L., and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657.

Hu, L., Fortin, N. J., and Ombao, H. (2019). Modeling high-dimensional multichannel brain signals. *Statistics in Biosciences*, 11(1):91–126.

Hu, L., Guindani, M., Fortin, N. J., and Ombao, H. (2020). A hierarchical bayesian model for differential connectivity in multi-trial brain signals. *Econometrics and Statistics*, 15:117–135.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., et al. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378.

Hyman, B. T., Van Hoesen, G. W., Damasio, A. R., and Barnes, C. L. (1984). Alzheimer's disease: cell-specific pathology isolates the hippocampal formation. *Science*, 225(4667):1168–1170.

Ibragimov, I. A. and Chernin, K. (1959). On the unimodality of geometric stable laws. *Theory of Probability & Its Applications*, 4(4):417–419.

Irie, K. (2019). Bayesian dynamic fused lasso. *arXiv preprint arXiv:1905.12275*.

Jacobs, P. A. and Lewis, P. A. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36.

Jentsch, C. and Reichmann, L. (2019). Generalized binary time series models. *Econometrics*, 7(4):47.

Jin, K., Peel, A. L., Mao, X. O., Xie, L., Cottrell, B. A., Henshall, D. C., and Greenberg, D. A. (2004). Increased hippocampal neurogenesis in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 101(1):343–347.

Jun, E., Na, K.-S., Kang, W., Lee, J., Suk, H.-I., and Ham, B.-J. (2020). Identifying resting-state effective connectivity abnormalities in drug-naïve major depressive disorder diagnosis via graph convolutional networks. *Human Brain Mapping*, 41(17):4997–5014.

Kalli, M. and Griffin, J. E. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793.

Kanter, M. (1975). Stable densities under change of scale and total variation inequalities. *The Annals of Probability*, pages 697–707.

Kapetanios, G., Marcellino, M., and Venditti, F. (2019). Large time-varying parameter vars: A nonparametric approach. *Journal of Applied Econometrics*, 34(7):1027–1049.

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76:408–423.

Kiebel, S. J., Klöppel, S., Weiskopf, N., and Friston, K. J. (2007). Dynamic causal modeling: a generative model of slice timing in fmri. *Neuroimage*, 34(4):1487–1496.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with arch models. *The review of economic studies*, 65(3):361–393.

Kleinhans, N. M., Richards, T., Sterling, L., Stegbauer, K. C., Mahurin, R., Johnson, L. C., Greenson, J., Dawson, G., and Aylward, E. (2008). Abnormal functional connectivity in autism spectrum disorders during face processing. *Brain*, 131(4):1000–1012.

Konovalov, A. and Krajbich, I. (2019). Over a decade of neuroeconomics: what have we learned? *Organizational Research Methods*, 22(1):148–173.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter vars. *Journal of Econometrics*, 177(2):185–198.

Koop, G. M. (2013). Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics*, 28(2):177–203.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019). Dynamic shrinkage processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):781–804.

Legramanti, S., Durante, D., and Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752.

Leistritz, L., Pester, B., Doering, A., Schiecke, K., Babiloni, F., Astolfi, L., and Witte, H. (2013). Time-variant partial directed coherence for analysing connectivity: a methodological study. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1997):20110616.

Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.-M., Schluep, M., Vuilleumier, P., and Van De Ville, D. (2013). Principal components of functional connectivity: a new approach to study dynamic brain connectivity during rest. *NeuroImage*, 83:937–950.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214.

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., and Coan, J. A. (2015). Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics*, 9(2):687 – 713.

Li, H., Wang, Y., Yan, G., Sun, Y., Tanabe, S., Liu, C.-C., Quigg, M., and Zhang, T. (2020). A bayesian state-space approach to mapping directional brain networks.

Liao, W., Ding, J., Marinazzo, D., Xu, Q., Wang, Z., Yuan, C., Zhang, Z., Lu, G., and Chen, H. (2011). Small-world directed networks in the human brain: multivariate granger causality analysis of resting-state fmri. *Neuroimage*, 54(4):2683–2694.

Lipster, R. and Shiryayev, A. (1972). Statistics of conditionally gaussian random sequences. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 389–422. University of California Press.

Liptser, R. S. and Shiryaev, A. N. (2013). *Statistics of random processes II: Applications*, volume 6. Springer Science & Business Media.

Litterman, R. B. et al. (1979). Techniques of forecasting using vector autoregressions. Technical report.

Liu, J., Ji, J., Xun, G., Yao, L., Huai, M., and Zhang, A. (2020). Ec-gan: Inferring brain effective connectivity via generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4852–4859.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*, volume 110. CRC Press.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite gaussian mixtures. *Statistics and Computing*, 26(1):303–324.

McCausland, W. J., Miller, S., and Pelletier, D. (2011). Simulation smoothing for state–space models: A computational efficiency analysis. *Computational Statistics & Data Analysis*, 55(1):199–212.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Monti, M. M. (2011). Statistical analysis of fmri time-series: a critical review of the glm approach. *Frontiers in human neuroscience*, 5:28.

Morf, M., Vieira, A., Kailath, T., et al. (1978). Covariance characterization by partial autocorrelation matrices. *The Annals of Statistics*, 6(3):643–648.

Nakajima, J. et al. (2011). Time-varying parameter var model with stochastic volatility: An overview of methodology and empirical applications. *Monetary and Economic Studies*, 29:107–142.

Neal, R. M. (2003). Slice sampling. *The annals of statistics*, 31(3):705–767.

Ombao, H., Fiecas, M., Ting, C.-M., and Low, Y. F. (2018). Statistical models for brain signals with properties that evolve across trials. *NeuroImage*, 180:609–618.

Omidvarnia, A., Azemi, G., Boashash, B., O'Toole, J. M., Colditz, P. B., and Vanhatalo, S. (2013). Measuring time-varying information flow in scalp eeg signals: orthogonalized partial directed coherence. *IEEE transactions on biomedical engineering*, 61(3):680–693.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2):425–449.

Ondrus, M., Olds, E., and Cribben, I. (2021). Factorized binary search: change point detection in the network structure of multivariate high-dimensional time series. *arXiv preprint arXiv:2103.06347*.

Pace-Schott, E. F. and Picchioni, D. (2017). Chapter 51 - neurobiology of dreaming. In Kryger, M., Roth, T., and Dement, W. C., editors, *Principles and Practice of Sleep Medicine (Sixth Edition)*, pages 529–538.e6. Elsevier, sixth edition edition.

Pankratz, A. (2012). *Forecasting with dynamic regression models*, volume 935. John Wiley & Sons.

Park, H.-J., Friston, K. J., Pae, C., Park, B., and Razi, A. (2018a). Dynamic effective connectivity in resting state fmri. *NeuroImage*, 180:594 – 608. Brain Connectivity Dynamics.

Park, H.-J., Friston, K. J., Pae, C., Park, B., and Razi, A. (2018b). Dynamic effective connectivity in resting state fmri. *Neuroimage*, 180:594–608.

Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer.

Phang, C.-R., Noman, F., Hussain, H., Ting, C.-M., and Ombao, H. (2019). A multi-domain connectome convolutional neural network for identifying schizophrenia from eeg connectivity patterns. *IEEE journal of biomedical and health informatics*, 24(5):1333–1343.

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002). Constructing first order stationary autoregressive models via latent processes. *Scandinavian Journal of Statistics*, 29(4):657–663.

Pitt, M. K. and Walker, S. G. (2005). Constructing stationary time series models using auxiliary variables with applications. *Journal of the American Statistical Association*, 100(470):554–564.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Prado, R. and West, M. (2010). *Time series: modeling, computation, and inference*. CRC Press.

Prieto, E., Eickmeier, S., and Marcellino, M. (2016). Time variation in macro-financial linkages. *Journal of Applied Econometrics*, 31(7):1215–1233.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252.

Razi, A. and Friston, K. J. (2016). The connected brain: causality, models, and intrinsic dynamics. *IEEE Signal Processing Magazine*, 33(3):14–35.

Rockova, V. and McAlinn, K. (2021). Dynamic variable selection with spike-and-slab process priors. *Bayesian Analysis*, 16(1):233–269.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Rowling, J. (2012). *Harry Potter and the Sorcerer's Stone*. Pottermore Limited.

Rue, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338.

Samdin, S. B., Ting, C.-M., Ombao, H., and Salleh, S.-H. (2016). A unified estimation framework for state-related changes in effective brain connectivity. *IEEE Transactions on Biomedical Engineering*, 64(4):844–858.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shahbaba, B., Li, L., Agostinelli, F., Saraf, M., Elias, G. A., Baldi, P., and Fortin, N. J. (2019). Hippocampal ensembles represent sequential relationships among discrete nonspatial events. *bioRxiv*, page 840199.

Shanbhag, D. N. and Sreehari, M. (1977). On certain self-decomposable distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 38(3):217–222.

Shephard, N. (1994). Partial non-gaussian state space. *Biometrika*, 81(1):115–131.

Shojaie, A. and Michailidis, G. (2010). Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics (Oxford, England)*, 26(18):i517–i523.

Sikka, A., Jamalabadi, H., Krylova, M., Alizadeh, S., van der Meer, J. N., Danyeli, L., Deliano, M., Vicheva, P., Hahn, T., Koenig, T., et al. (2020). Investigating the temporal dynamics of electroencephalogram (eeg) microstates using recurrent neural networks. *Human brain mapping*, 41(9):2334–2346.

Simon, T. (2011). Multiplicative strong unimodality for positive stable laws. *Proceedings of the American Mathematical Society*, 139(7):2587–2595.

Sims, C. A. (1993). A nine-variable probabilistic macroeconomic forecasting model. In *Business cycles, indicators, and forecasting*, pages 179–212. University of Chicago press.

Stephan, K. E. and Friston, K. J. (2010). Analyzing effective connectivity with functional magnetic resonance imaging. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(3):446–459.

Stephan, K. E., Penny, W. D., Moran, R. J., den Ouden, H. E., Daunizeau, J., and Friston, K. J. (2010). Ten simple rules for dynamic causal modeling. *Neuroimage*, 49(4):3099–3109.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39:3–33.

Sun, W. W. and Li, L. (2019). Dynamic tensor clustering. *Journal of the American Statistical Association*, pages 1–28.

Taghia, J., Ryali, S., Chen, T., Supekar, K., Cai, W., and Menon, V. (2017). Bayesian switching factor analysis for estimating time-varying functional connectivity in fmri. *Neuroimage*, 155:271–290.

TRACY, J. I. and BOSWELL, S. B. (2008). Mesial temporal lobe epilepsy: a model for understanding the relationship between language and memory. In *Handbook of the neuroscience of language*, pages 319–328. Elsevier.

Trujillo-Barreto, N. J., Aubert-Vázquez, E., and Penny, W. D. (2008). Bayesian m/eeg source reconstruction with spatio-temporal priors. *Neuroimage*, 39(1):318–335.

Trujillo-Barreto, N. J., Aubert-Vázquez, E., and Valdés-Sosa, P. A. (2004). Bayesian model averaging in eeg/meg imaging. *NeuroImage*, 21(4):1300–1319.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289.

Ullman, J. B. and Bentler, P. M. (2003). Structural equation modeling. *Handbook of psychology*, pages 607–634.

van de Ven, V. G., Formisano, E., Prvulovic, D., Roeder, C. H., and Linden, D. E. (2004). Functional connectivity as revealed by spatial independent component analysis of fmri measurements during rest. *Human brain mapping*, 22(3):165–178.

Velu, R. P., Reinsel, G. C., and Wichern, D. W. (1986). Reduced rank models for multiple time series. *Biometrika*, 73(1):105–118.

Vidaurre, D., Smith, S. M., and Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, D., Zheng, Y., Lian, H., and Li, G. (2021). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, pages 1–19.

Wang, Y. and Ho, H. (2016). Statistical analysis of electroencephalograms. *Handbook of Neuroimaging Data Analysis*, pages 523–565.

Wang, Y., Ting, C.-M., and Ombao, H. (2016). Modeling effective connectivity in high-dimensional cortical source signals. *IEEE Journal of Selected Topics in Signal Processing*, 10(7):1315–1325.

Warnick, R., Guindani, M., Erhardt, E., Allen, E., Calhoun, V., and Vannucci, M. (2018). A bayesian approach for estimating dynamic functional network connectivity in fmri data. *Journal of the American Statistical Association*, 113(521):134–151.

Welvaert, M. and Rosseel, Y. (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for fmri data. *PloS one*, 8(11):e77089.

West, M. J., Coleman, P. D., Flood, D. G., and Troncoso, J. C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and alzheimer's disease. *The Lancet*, 344(8925):769–772.

Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50(1-2):129–134.

Xiong, X. and Cribben, I. (2021). Beyond linear dynamic functional connectivity: a vine copula change point model. *bioRxiv*.

Zarghami, T. S. and Friston, K. J. (2020a). Dynamic effective connectivity. *NeuroImage*, 207:116453.

Zarghami, T. S. and Friston, K. J. (2020b). Dynamic effective connectivity. *Neuroimage*, 207:116453.

Zevin, J. (2009). Word recognition. In Squire, L. R., editor, *Encyclopedia of Neuroscience*, pages 517–522. Academic Press, Oxford.

Zhang, L., Guindani, M., and Vannucci, M. (2015a). Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):21–41.

Zhang, L., Guindani, M., and Vannucci, M. (2015b). Bayesian models for functional magnetic resonance imaging data analysis. *WIREs Computational Statistics*, 7:21–41.

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). A spatio-temporal nonparametric bayesian variable selection model of fmri data for clustering correlated time courses. *NeuroImage*, 95:162–175.

Zhang, W., Cribben, I., Petrone, S., and Guindani, M. (2021). Bayesian time-varying tensor vector autoregressive models for dynamic effective connectivity. *arXiv preprint arXiv:2106.14083*.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.