Markov Switching Multiple-equation Tensor Regressions

Roberto Casarin[a], Radu V. Craiu[b,*], Qing Wang[a]

[a] *Ca' Foscari University of Venice, Italy*
[b] *University of Toronto, Canada*

Abstract

A new flexible tensor model for multiple-equation regressions that accounts for latent regime changes is proposed. The model allows for dynamic coefficients and multi-dimensional covariates that vary across equations. The coefficients are driven by a common hidden Markov process that addresses structural breaks to enhance the model flexibility and preserve parsimony. A new soft PARAFAC hierarchical prior is introduced to achieve dimensionality reduction while preserving the structural information of the covariate tensor. The proposed prior includes a new multi-way shrinking effect to address over-parametrization issues while preserving interpretability and model tractability. Theoretical results are derived to help with the choice of the hyperparameters. An efficient Markov chain Monte Carlo (MCMC) algorithm based on random scan Gibbs and back-fitting strategy is designed with priority placed on computational scalability of the posterior sampling. The validity of the MCMC algorithm is demonstrated theoretically, and its computational efficiency is studied using numerical experiments in different parameter settings. The effectiveness of the model framework is illustrated using two original real data analyses. The proposed model exhibits superior performance compared to the current benchmark, Lasso regression.

Keywords: Bayesian inference, Dimensionality reduction, Markov switching, MCMC, Tensor regression.
2020 MSC: Primary 62H12, Secondary 62F15, 62P20

1. Introduction

As data grow in volume and complexity, it is increasingly common to record them as high-dimensional arrays or tensors. Such structures appear in many applications and fields, such as neuroimaging [19, 37], biostatistics, financial networks [7], or even more generally, in time series [8]. People are often interested in characterizing the relationship between a tensor predictor and a scalar outcome [19] or tensor outcome [38]. Tensor regression has been studied extensively in a linear model framework. Nevertheless, a common challenge within the framework of regression is model misspecification. One of the sources of model misspecification is the presence of dynamic regimes, which naturally call for models with time-varying parameters.

In this paper, we assume a Hidden Markov chain dynamics for the regression coefficients because it allows for model parsimony while preserving a high level of flexibility compared to other time-varying parameter models, which typically require a larger number of factors or parameters [e.g., see 24, 28]. Hidden Markov Models (HMM), also known as Markov Switching (MS), have been introduced to capture structural changes and regimes [17]. Since the seminal works on univariate autoregressive HMMs [14, 22], HMMs have been extended in different directions. The univariate extensions include the MS ARMA [10], MS stochastic volatility models [36], MS with time-varying transition [25] and random transition [4, 6]. The multivariate extensions include MS Vector Autoregressive (VAR) models introduced by [35], MS stochastic volatility VARs [13], MS graphical VARs [5] and MS panel data models [1, 9, 12]. Efforts have been made to address abrupt structural changes in temporal networks by [7]. They proposed a tensor-on-tensor logistic regression model combining a low-rank decomposition and HMM to model the coefficient tensor. The contribution of our paper is multifold. First, we extend the soft tensor linear regression models [19, 32, 38] to an HMM (or MS) framework to accommodate structural breaks. We assume the margins in the Parallel Factor (PARAFAC) representation, also called CANDECOMP/PARAFAC or Polyadic Decomposition [e.g., see 21, 26], of the coefficient tensor are driven by a common hidden Markov chain process. Thus, a flexible time-varying parameter model is obtained with a small increase in the latent space dimension and the number of parameters. Second, we consider a multi-equation setting in which the latent process provides a time-varying structure to several tensor regression models involving different response variables and, possibly, different sets of covariates. Third, we propose a

[*] Corresponding author at: University of Toronto, Canada. Email addresses: r.casarin@unive.it (R. Casarin), radu.craiu@utoronto.ca (R.V. Craiu), qing.wang@unive.it (Q. Wang)

Bayesian inference procedure that relies on numerical exploration of the posterior via a new and efficient Gibbs sampler, which reduces computational costs and improves scalability. Finally, using a common latent process is intended to address two goals: 1) it facilitates the integration of information about the latent process from multiple outcomes, and 2) it robustifies the estimation of regime changes, which must be supported by multiple outcome variables simultaneously.

The complex structure of multi-dimensional data naturally poses challenges such as over-parame-trization and overfitting issues. A simple approach in the tensor regression framework is to vectorize the tensor predictor and regress the response variable on a large vector of tensor entries with some form of penalization and variable selection. However, this approach completely ignores the structural relationships embedded in the tensor predictor. Most research on tensor regression focuses on dimensionality reduction for tensor predictors or coefficients. Various dimensionality reduction strategies have been proposed to cope with these issues. For instance, [44] adopted a two-stage procedure to study the relationship between individuals' structural connectomes and human traits, using principal component analysis on the tensor predictors to achieve dimension reduction and then fitting a model using lower dimensional summaries of tensor predictors. Similarly, [11] carried out SVD on high dimensional fMRI data to study the relationship between functional connectivity and Alzheimer's disease risk. However, this approach suffers from the unsupervised nature of PCA, and the loss of structural information on the tensor predictors and interpretation of estimated coefficients could be difficult. Thus, we follow a different approach based on the reduction of dimensionality of tensor coefficients, which preserves the structural dependence of the predictor tensor.

Within the frequentist paradigm, [43] applied Tucker decomposition on the tensor coefficients and proposed a fast algorithm (Tensor Projected Gradient) to minimize the empirical loss function. [46] used PARAFAC decomposition, a special case of Tucker decomposition, on the tensor coefficients and relied on maximum likelihood estimation to perform neuroimaging data analysis. [27] used a neural network combined with Tucker representation to address multi-way data analysis. In this paper, we follow a flexible Bayesian modelling approach.

Within the Bayesian paradigm, [42, 45] proposed non-parametric methods based on Gaussian Process priors. In a scalar on tensor regression framework, [20] proposed a novel multi-way shrinkage prior on the PARAFAC representation of the coefficient tensor. Their work was extended by [37], who explored a more general Tucker decomposition on the tensor coefficients. In follow-up work, [19] proposed a Bayesian network shrinkage prior and used a spike-and-slab prior distribution to determine which brain nodes are most influential to creativity. In the case of tensor on tensor regression, [38] proposed to use the Tucker decomposition of the coefficient tensor without assuming the dimension of the core tensor. In this paper, we follow a soft PARAFAC framework [32] where the hierarchical prior distribution of [20] is modified to allow the coefficient tensor to deviate randomly from the rigid low-rank PARAFAC representation. We modified the multi-way shrinkage priors from [32] and [20] to facilitate prior calibration and to improve the tractability of the conditional posterior distributions. We developed an efficient Markov chain Monte Carlo (MCMC) algorithm to achieve better scalability, relying on a random scan Gibbs sampler within the back-fitting strategy, usually employed in Bayesian high-dimensional models [23, 29, 47].

The paper is organized as follows: in Section 2, we revisit the concept of soft PARAFAC decomposition for dimensionality reduction and introduce the Markov-Switching multiple-equation Tensor Regression (MSTR) and the Bayesian framework for inference. In Section 3, we propose a new MCMC algorithm based on Random Partial Scan Gibbs and back–fitting strategy, prove its ergodicity and demonstrate its performance using numerical experiments (simulation results are shown in Appendix C of the Supplement). In Section 4, we test our model with two applications that show the gain in performance in terms of in-sample fitting and out-of-sample forecasting. The paper ends with Section 5, which contains conclusions and future promising directions.

2. A Markov-switching multiple-equation tensor regression model

To simplify the exposition in this section, we assume covariates are common to all the equations. Furthermore, the error terms are assumed to be independent across equations, but the approach generalizes to equation-specific covariate tensors and dependent errors. Our MSTR model assumes a system of N equations with time-varying parameters

$$\begin{cases} y_{1,t} = \mu_1(s_t) + \langle B_1(s_t), X_t \rangle + \sigma_1(s_t)\varepsilon_{1,t}, \\ \quad\vdots \\ y_{N,t} = \mu_N(s_t) + \langle B_N(s_t), X_t \rangle + \sigma_N(s_t)\varepsilon_{N,t}, \end{cases} \tag{1}$$

$t \in \{1, 2, \dots, T\}$, where $y_{\ell,t}$, $\ell \in \{1, \dots, N\}$ are scalar response variables, X_t is a $p_1 \times \cdots \times p_M$ covariate tensor, $B_\ell(s_t)$, $\ell \in \{1, \dots, N\}$ are $p_1 \times \cdots \times p_M$ coefficient tensors, with M denoting the number of tensor modes, $\varepsilon_{\ell,t}$, $\ell \in \{1, \dots, N\}$ are i.i.d. from $\mathcal{N}(0, 1)$, $\{s_t, t \in \{1, \dots, T\}\}$ is a common latent process, and $\langle \cdot, \cdot \rangle$ denotes the

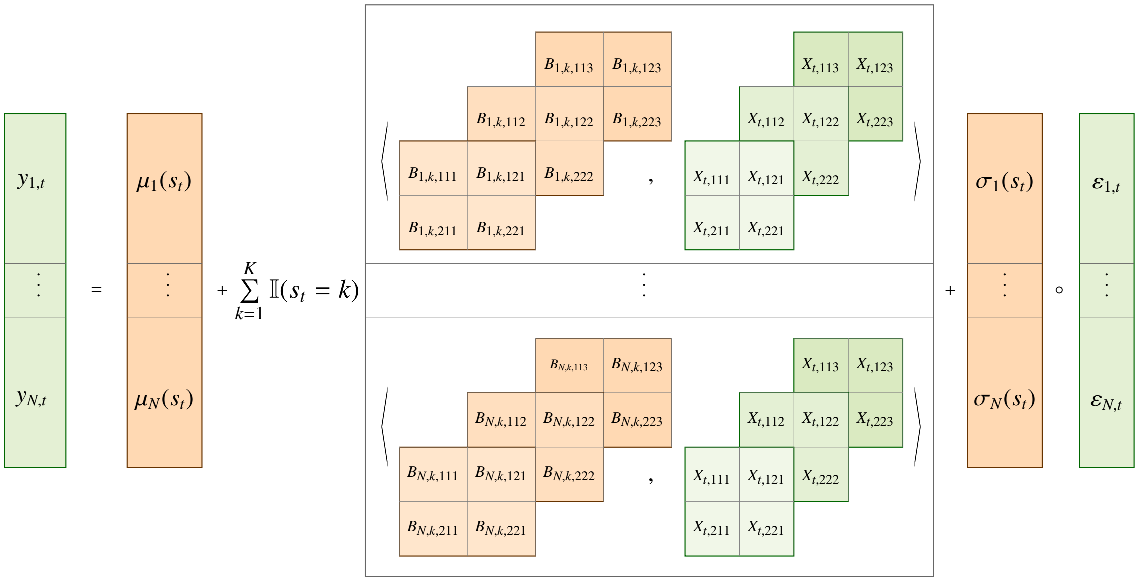
$y_{1,t}$
$y_{N,t}$
=
$\mu_1(s_t)$
$\mu_N(s_t)$
$+\sum_{k=1}^{K}\mathbb{I}(s_t=k)$
$B_{1,k,113}$ $B_{1,k,123}$ $B_{1,k,112}$ $B_{1,k,122}$ $B_{1,k,223}$ $B_{1,k,111}$ $B_{1,k,121}$ $B_{1,k,222}$ $B_{1,k,211}$ $B_{1,k,221}$
$B_{N,k,113}$ $B_{N,k,123}$ $B_{N,k,112}$ $B_{N,k,122}$ $B_{N,k,223}$ $B_{N,k,111}$ $B_{N,k,121}$ $B_{N,k,222}$ $B_{N,k,211}$ $B_{N,k,221}$
$X_{t,113}$ $X_{t,123}$ $X_{t,112}$ $X_{t,122}$ $X_{t,223}$ $X_{t,111}$ $X_{t,121}$ $X_{t,222}$ $X_{t,211}$ $X_{t,221}$
+
$\sigma_1(s_t)$
$\sigma_N(s_t)$
$\circ$
$\varepsilon_{1,t}$
$\varepsilon_{N,t}$

Fig. 1: Graphic representation of a multiple-equation tensor regression with switching parameters $\mu_\ell(s_t)$, $B_\ell(s_t) \in \mathbb{R}^{2\times2\times3}$ and $\sigma_\ell(s_t)$, and covariate tensor $X_t \in \mathbb{R}^{2\times2\times3}$. Green shades denote response variables and covariates, and orange shades denote parameters and latent variables.

inner product for tensors [21, 26]. While the dependent variables are conditionally independent, joint inference remains essential in our tensor-on-tensor regression because the equations are governed by a common latent process which captures the underlying dynamics across variables.

The latent process is a K-state homogeneous Markov chain with transition probability $\Pr(s_t = j|s_{t-1} = i) = p_{i,j}$, $i, j \in \{1, \dots, K\}$ and the tensor regression parametrization used is

$$\mu_\ell(s_t) = \sum_{k=1}^{K} \mu_{\ell,k}\mathbb{I}(s_t = k), \quad B_\ell(s_t) = \sum_{k=1}^{K} B_{\ell,k}\mathbb{I}(s_t = k), \quad \sigma^2_\ell(s_t) = \sum_{k=1}^{K} \sigma^2_{\ell,k}\mathbb{I}(s_t = k). \tag{2}$$

Alternative parameterizations for the coefficients can be used; for example, see [17] for conditionally linear single-equation models and [13] for conditionally linear multiple-equation models. A graphical representation of our tensor regression model with N equations and a 3-mode covariate tensor of dimension $p_1 = 2$, $p_2 = 2$ and $p_3 = 3$ is shown in Fig. 1.

Since, in many applications, the number of covariates in (1) is large, a dimensionality reduction strategy is needed. In this paper, we follow [32] and [7, 8] and consider a low-rank representation combined with a hierarchical prior distribution. The hierarchical prior allows for shrinking effects in the coefficient tensors $B_{\ell,k}$, $k \in \{1, \dots, K\}$, and the low-rank representation induces further shrinking effects along different modes. We assume a PARAFAC representation and decompose the state-specific coefficient tensor as follows

$$B_{\ell,k} = \sum_{d=1}^{D} B^{(d)}_{\ell,1,k} \circ B^{(d)}_{\ell,2,k} \circ \cdots \circ B^{(d)}_{\ell,M,k}, \tag{3}$$

where $\circ$ is the element-by-element Hadamard product, $B^{(d)}_{\ell,m,k}$, $m \in \{1, 2, \dots, M\}$ are multiplicative factors ([32]).

The hierarchical prior distribution includes three stages. At the first stage, an inverse gamma prior distribution $\mathcal{IG}(a_\sigma, b_\sigma)$ with shape and scale parameters a_σ and b_σ is assumed for $\sigma^2_{\ell,k}$ and a tensor-variate normal distribution [30, 31] is assumed for the coefficient tensor

$$B^{(d)}_{\ell,m,k} \sim \mathcal{TN}_{p_1,\dots,p_M}\left(G^{(d)}_{\ell,m,k}, \tau_{\ell,k}\kappa^2_{\ell,m,k}\zeta^{(d)}_{\ell,k} I_{p_1}, \dots, I_{p_M}\right), \tag{4}$$

$\ell \in \{1, \dots, N\}$, $m \in \{1, \dots, M\}$, $d \in \{1, \dots, D\}$, $k \in \{1, \dots, K\}$, where $\mathcal{TN}_{p_1,\dots,p_M}(G, U_1, \dots, U_M)$ denotes the tensor-variate normal distribution with $p_1 \times \cdots \times p_M$ location tensor G, $p_m \times p_m$ covariance matrix U_m for the mth mode elements. The location $G^{(d)}_{\ell,m,k}$ is parametrized as follows:

$$G^{(d)}_{\ell,m,k} = \boldsymbol{\iota}_{p_1} \otimes \cdots \otimes \boldsymbol{\iota}_{p_{m-1}} \otimes \boldsymbol{\gamma}^{(d)}_{\ell,m,k} \otimes \boldsymbol{\iota}_{p_{m+1}} \otimes \cdots \otimes \boldsymbol{\iota}_{p_M}, \tag{5}$$

where $\otimes$ denotes the outer product, $\boldsymbol{\iota}_n = (1, \dots, 1)^\top$ is the n-dimensional unit vector, $\boldsymbol{\gamma}^{(d)}_{\ell,m,k}$ is the mth PARAFAC margins of size p_m.

In the conditional mean of the components $B^{(d)}_{\ell,k}$ given $\boldsymbol{\gamma}_{\ell,k} = \{(\boldsymbol{\gamma}^{(d)}_{\ell,1,k}, \dots, \boldsymbol{\gamma}^{(d)}_{\ell,M,k}), d \in \{1, \dots, D\}\}$, we have

$$\begin{aligned} \mathrm{E}\left(B_{\ell,k} \mid \boldsymbol{\gamma}_{\ell,k}\right) &= \sum_{d=1}^{D} \mathrm{E}\left(B^{(d)}_{\ell,1,k} \mid \boldsymbol{\gamma}^{(d)}_{\ell,1,k}\right) \circ \cdots \circ \mathrm{E}\left(B^{(d)}_{\ell,M,k} \mid \boldsymbol{\gamma}^{(d)}_{\ell,M,k}\right) \\ &= \sum_{d=1}^{D} \left(G^{(d)}_{\ell,1,k} \circ \cdots \circ G^{(d)}_{\ell,M,k}\right) = \sum_{d=1}^{D} \boldsymbol{\gamma}^{(d)}_{\ell,1,k} \otimes \cdots \otimes \boldsymbol{\gamma}^{(d)}_{\ell,M,k}. \end{aligned} \tag{6}$$

In the second stage, we assume that the margins from the PARAFAC decomposition are independent and follow multivariate normal distributions

$$\boldsymbol{\gamma}^{(d)}_{\ell,m,k} \sim \mathcal{N}_{p_m}(\mathbf{0}, \tau_{\ell,k}\zeta^{(d)}_{\ell,k} W^{(d)}_{\ell,m,k}), \tag{7}$$

and assume the distributions are centered around the null vector with scale given by the product of the scalars $\tau_{\ell,k}$ and $\zeta^{(d)}_{\ell,k}$ and the diagonal matrix $W^{(d)}_{\ell,m,k} = \mathrm{diag}(w^{(d)}_{\ell,m,1,k}, \dots, w^{(d)}_{\ell,m,j_m,k}, \dots, w^{(d)}_{\ell,m,p_m,k})$. This random scale specification allows for shrinkage at different levels.

At the third stage of the prior, we borrow from [20] and specify the scale prior distributions to induce shrinkage across components and rows

$$\begin{aligned} \tau_{\ell,k} &\sim \mathcal{G}a(a_\tau, b_\tau), \quad \kappa^2_{\ell,m,k} \sim \mathcal{G}a(a_\kappa, b_\kappa), \quad w^{(d)}_{\ell,m,j_m,k} \sim \mathcal{E}xp((\lambda^{(d)}_{\ell,m,k})^2/2),\ j_m \in \{1, \dots, p_m\} \\ \lambda^{(d)}_{\ell,m,k} &\sim \mathcal{G}a(a_\lambda, b_\lambda), \quad \left(\zeta^{(1)}_{\ell,k}, \dots, \zeta^{(D)}_{\ell,k}\right) \sim \mathcal{D}ir(\alpha/D, \dots, \alpha/D), \end{aligned}$$

where $\mathcal{G}a(a, b)$, $\mathcal{E}xp(\lambda)$ and $\mathcal{D}ir(\nu_1, \dots, \nu_D)$ denote the Gamma, Exponential and Dirichlet distributions, respectively. Compared to [32] our prior assumes the global scale $\tau_{\ell,k}$ contributes not only to the variance of $\boldsymbol{\gamma}^{(d)}_{\ell,m,k}$ but also to that of one of the tensor coefficients $B^{(d)}_{\ell,m,k}$. This allows for stronger shrinkage effects and a full factorization of the prior variance, as detailed below. The scale parameter $w^{(d)}_{\ell,m,j_m,k}$ is the j_mth element of the diagonal of $W^{(d)}_{\ell,m,k}$, with $j_m \in \{1, \dots, p_m\}$. It is a row-specific parameter that shrinks the individual elements, $\gamma^{(d)}_{\ell,m,j_m,k}$, of the PARAFAC margins $\boldsymbol{\gamma}^{(d)}_{\ell,m,k}$. Together with the prior on $\lambda^{(d)}_{\ell,m,k}$, they lead to an adaptive LASSO-type penalty on $\boldsymbol{\gamma}^{(d)}_{\ell,m,k}$ [2]. The parameters $\zeta^{(d)}_{\ell,k}$ are component-specific and allow a subset of the D components to contribute substantially to the PARAFAC approximation while leaving the values of other components close to zero. The transition probabilities $(p_{1k}, \dots, p_{Kk})$ are assumed to follow a Dirichlet distribution

$$(p_{1,k}, \dots, p_{K,k}) \sim \mathcal{D}ir(\nu_1, \dots, \nu_K), \tag{8}$$

for $k \in \{1, \dots, K\}$. The choice of the prior hyperparameter value is crucial in Bayesian inference and can greatly affect the model's performance. We turn to study the induced prior for the ℓth coefficient tensor $B_{\ell,k}$ to elicit the default choice of hyperparameters. In particular, using a multiple-index notation, the variance of the $\tilde{j}$th entry $B_{\ell,k,\tilde{j}}$ of coefficient tensor $B_{\ell,k}$ for the soft PARAFAC, with $\tilde{j} = (j_1, \dots, j_M)$, can be written as a function of the hyperparameters:

$$\mathrm{V}(B_{\ell,k,\tilde{j}}) = C_\tau C_\zeta \left(\frac{a_\kappa}{b_\kappa} + \frac{2b^2_\lambda}{(a_\lambda - 1)(a_\lambda - 2)}\right)^2, \quad C_\zeta = \frac{\frac{\alpha}{D} + 1}{\alpha + 1}, \quad C_\tau = \frac{a_\tau(a_\tau + 1)}{b^2_\tau}. \tag{9}$$

Moreover, the variance of the coefficient entries for the hard PARAFAC is:

$$\mathrm{V}^{\text{hard}}\left(B_{\ell,k,\tilde{j}}\right) = C_\tau C_\zeta \left(\frac{2b^2_\lambda}{(a_\lambda - 1)(a_\lambda - 2)}\right)^2.$$

Thanks to the assumptions on the global shrinkage scale, the expression for the prior variances in soft and hard PARAFAC specifications factorize in a global shrinking effect function of a_τ, b_τ and local shrinking effects. The ratio between the two variances shows that the difference between soft and hard PARAFAC is not affected by the global shrinkage scale, thus providing better interpretability of the hyperparameters than [20] while preserving the tractability of the full conditional distributions in the Gibbs sampling procedure used for posterior approximation. We define the relative additional variance (AV) introduced by the softening of the PARAFAC as follows:

$$AV = \frac{\mathrm{V}\left(B_{\ell,k,\tilde{j}}\right) - \mathrm{V}^{\text{hard}}\left(B_{\ell,k,\tilde{j}}\right)}{\mathrm{V}\left(B_{\ell,k,\tilde{j}}\right)} = 1 - \left(1 + \frac{a_\kappa}{b_\kappa}\frac{(a_\lambda - 1)(a_\lambda - 2)}{2b^2_\lambda}\right)^{-2},$$

which depends on the hyper-parameters of the local shrinkage scales and can be used to elicit the values of the hyper-parameters.

Proposition 1. *For a tensor coefficient, target variance $V^* \in (0, \infty)$, target additional variance $AV^* \in [0, 1)$, we have the following expression,*

$$\frac{a_\kappa}{b_\kappa} = \frac{b_\tau}{a_\tau}\sqrt{\frac{a_\tau V^*}{(a_\tau + 1)C_\zeta}}\left(1 - \sqrt{1 - AV^*}\right). \tag{10}$$

Proposition 1 and the identities in (9) are used in the simulations and empirical applications to help choose hyperparameters. In particular, we impose restrictions on the induced prior variance such that $\mathrm{V}(B_{\ell,k,\tilde{j}}) = 1$ and $AV = 10\%$. Moreover, we set $\alpha = 1, a_\tau = 3, a_\kappa = 0.5, a_\lambda = 3, b_\lambda = \sqrt[2M]{a_\lambda}$ following [32] and we compute the values of b_τ and b_κ from (9) and (10) for which $V^* = 1$ and $AV^* = 10\%$.

The choice of rank D for the soft PARAFAC decomposition of the tensor coefficient can lead to significant changes in computational costs, with a higher value of D triggering drastic increases in computational time. However, the increase in D doesn't necessarily guarantee a vast boost in inferential performance. Intuitively, the soft PARAFAC can expand away from the low-rank hard PARAFAC structure and achieve a higher-rank representation of the tensor coefficient. We will provide a discussion in the next section.

3. Posterior approximation

The joint posterior distribution is not tractable, so we develop an MCMC algorithm to sample from it. Specifically, we use a Gibbs sampling procedure which combines two sampling strategies: i) back-fitting sampling [23] for the coefficients and ii) forward filtering and backward sampling for the latent states [17]. To cope with the computational cost of the Monte Carlo approximation, we implement a version of the random scan Gibbs [47].

3.1. Back-fitting representation

In this section, we assume covariate and coefficient tensors of general order. Let us denote with $B^{(d)}_{\ell,m,\tilde{j}_m,k}$ the $p_1 \times \cdots \times p_{m-1} \times p_{m+1} \times \cdots \times p_M$ tensor with $M - 1$ modes which is the j_mth slice of tensor $B^{(d)}_{\ell,m,k}$ along the mode m in the regime k for the ℓth equation, where $\tilde{j}_m = \{(j_1, \ldots, j_m, \ldots, j_M), j_h \in \{1, \ldots, p_h\}, \forall h \neq m\}$ is the collection of index values along $M - 1$ modes while keeping fix the index j_m of the mode m.

For $\ell \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$, $m \in \{1, \ldots, M\}$, $d \in \{1, \ldots, D\}$ and $j_m \in \{1, \ldots, p_m\}$ define the $q_m \times 1$ vector $\boldsymbol{\beta}^{(d)}_{\ell,m,j_m,k} = \mathrm{vec}(B^{(d)}_{\ell,m,\tilde{j}_m,k})$, with $q_m = \prod_{l \neq m} p_l$, obtained by stacking vertically all 1-mode fibers of the tensor following a lexicographic order of the indexes [21, 26]. We further define the collections $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{1,k}, \ldots, \boldsymbol{\beta}_{N,k})$ and $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{1,k}, \ldots, \boldsymbol{\gamma}_{N,k})$, with $\boldsymbol{\beta}_{\ell,k} = (\boldsymbol{\beta}^{(1)}_{\ell,1,1,k}, \ldots, \boldsymbol{\beta}^{(d)}_{\ell,m,j_m,k}, \ldots, \boldsymbol{\beta}^{(D)}_{\ell,M,p_M,k})$ and $\boldsymbol{\gamma}_{\ell,k} = (\gamma^{(1)}_{\ell,1,1,k}, \ldots, \gamma^{(d)}_{\ell,m,j_m,k}, \ldots, \gamma^{(D)}_{\ell,M,p_M,k})^\top$, where $\gamma^{(d)}_{\ell,m,j_m,k}$ is the j_mth entry, $j_m = 1, \ldots, p_m$, of the PARAFAC margin $\boldsymbol{\gamma}^{(d)}_{\ell,m,k}$ defined in Section 2.

We summarize our Bayesian model in the Directed Acyclic Graph (DAG) representation of Fig. 2 where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{N,t})^\top$ is the collection of response variables across equations, $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_K)$ is the collection of transition probabilities, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$, $\boldsymbol{\sigma}^2 = (\boldsymbol{\sigma}^2_1, \ldots, \boldsymbol{\sigma}^2_K)$, and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ denote the collections across equations and states of the regression coefficients, the PARFAC components, the error scale parameters and intercepts, respectively, where $\boldsymbol{\mu}_k = (\mu_{1,k}, \ldots, \mu_{N,k})^\top$, $\boldsymbol{\sigma}^2_k = (\sigma^2_{1,k}, \ldots, \sigma^2_{N,k})^\top$ and $\mathbf{p}_k = (p_{1,k}, \ldots, p_{K,k})^\top$. In the same DAG $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)^\top$, $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$, and $\boldsymbol{\kappa}^2 = (\boldsymbol{\kappa}^2_1, \ldots, \boldsymbol{\kappa}^2_K)$ denote the collections of the hyper-parameters at the second and third stage of the hierarchical prior where $\boldsymbol{\zeta}_k = (\boldsymbol{\zeta}_{1,k}, \ldots, \boldsymbol{\zeta}_{N,k})$, $\boldsymbol{\zeta}_{\ell,k} = (\zeta^{(1)}_{\ell,k}, \ldots, \zeta^{(d)}_{\ell,k}, \ldots, \zeta^{(D)}_{\ell,k})^\top$, $\boldsymbol{\lambda}_k = (\lambda_{1,k}, \ldots, \lambda_{N,k})$, $\boldsymbol{\lambda}_{\ell,k} = (\lambda^{(1)}_{\ell,1,k}, \ldots, \lambda^{(d)}_{\ell,m,k}, \ldots, \lambda^{(D)}_{\ell,M,k})^\top$, $\boldsymbol{w}_k = (\boldsymbol{w}_{1,k}, \ldots, \boldsymbol{w}_{N,k})$, $\boldsymbol{w}_{\ell,k} = (w^{(1)}_{\ell,1,1,k}, \ldots, w^{(d)}_{\ell,m,j_m,k}, \ldots, w^{(D)}_{\ell,M,p_M,k})^\top$, and $\boldsymbol{\kappa}^2_k = (\boldsymbol{\kappa}^2_{1,k}, \ldots, \boldsymbol{\kappa}^2_{N,k})$, $\boldsymbol{\kappa}^2_{\ell,k} = (\kappa^2_{\ell,1,k}, \ldots, \kappa^2_{\ell,M,k})^\top$.

The MCMC sampler proposed in the next section relies on the following equivalent representation of the MSTR model.

Proposition 2. *The model in* (1) *can be written as:*

$$y_{\ell,t} = \boldsymbol{\beta}^{(d)}_{\ell,m,j_m}(s_t)^\top \Psi^{(d)}_{\ell,m,j_m,t}(s_t) + R^{(d)}_{\ell,m,j_m,t}(s_t) + R^{(d)}_{\ell,t}(s_t) + \sigma^2_\ell(s_t)\varepsilon_{\ell,t},$$

$\ell \in \{1, \ldots, N\}$, *where the residual terms* $R^{(d)}_{\ell,t}(s_t)$ *and* $R^{(d)}_{\ell,m,j_m,t}(s_t)$ *and the auxiliary covariate vector* $\Psi^{(d)}_{\ell,m,j_m,t}(s_t)$ *are defined as follows:*

$$R^{(d)}_{\ell,t}(s_t) = \sum_{d' \neq d} \left\langle B^{(d')}_{\ell,1}(s_t) \circ \cdots \circ B^{(d')}_{\ell,M}(s_t), X_t \right\rangle, \quad R^{(d)}_{\ell,m,j_m,t}(s_t) = \left\langle (B^{(d)}_{\ell,1}(s_t) \circ \cdots \circ B^{(d)}_{\ell,M}(s_t))_{-j_m}, (X_t)_{-j_m} \right\rangle,$$

$$\Psi^{(d)}_{\ell,m,j_m,t}(s_t) = vec\left((B^{(d)}_{\ell,1}(s_t) \circ \cdots \circ B^{(d)}_{\ell,m-1}(s_t) \circ B^{(d)}_{\ell,m+1}(s_t) \circ \cdots \circ B^{(d)}_{\ell,M}(s_t) \circ X_t)_{\tilde{j}_m}\right)$$

with $B^{(d)}_{\ell,m}(s_t) = \sum_{k=1}^K B^{(d)}_{\ell,m,k}\mathbb{I}(s_t = k)$ *a Markov-switching tensor coefficient and* $(A)_{-j_m}$ *the tensor obtained removing from A the* j_m*th slice along the mode m.*

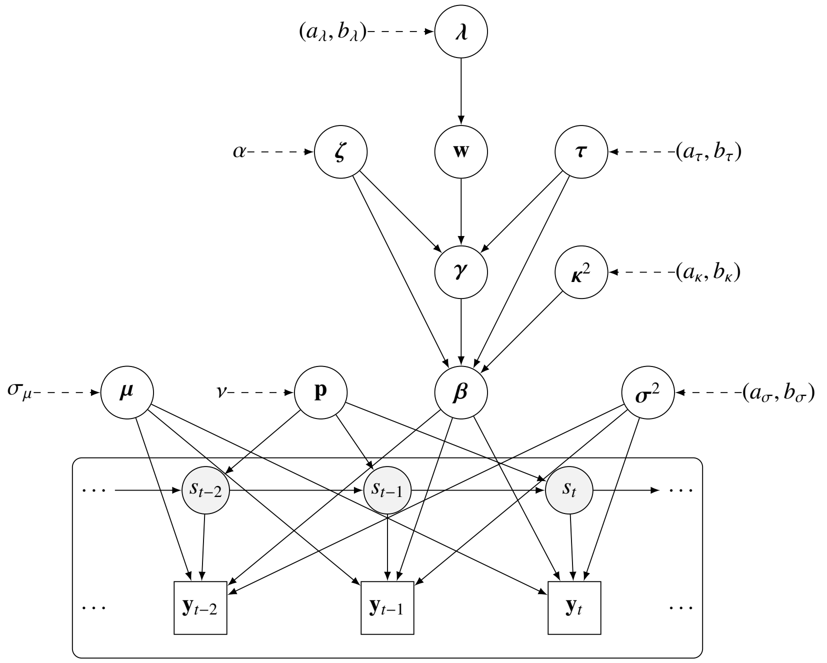

Fig. 2: Directed Acyclic Graph of the Bayesian Markov-switching Tensor Regression model. It exhibits the hierarchical structure of the observations $\mathbf{y}_t$ (boxes), the latent state variables s_t (grey circles), the parameters $\boldsymbol{\beta}^{(d)}_{\ell,m,j_m,k}$, $\mu_{\ell,k}$, $\mathbf{p}_k$ and $\sigma^2_{\ell,k}$, the hyper-parameters of the first stage $\gamma^{(d)}_{\ell,m,j_m,k}$ and $\kappa^2_{\ell,m,k}$, the second stage $\tau_{\ell,k}$, $\zeta^{(d)}_{\ell,m,k}$ and $w^{(d)}_{\ell,m,j_m,k}$ and the third stage $\lambda^{(d)}_{\ell,m,k}$ (white circles). The directed arrows show the conditional independence structure of the model.

3.2. Sampling method

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ be the collection of the state-specific parameters $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}^2_k, \mathbf{p}_k, \tau_k, \boldsymbol{\zeta}_k, \boldsymbol{w}_k, \lambda_k, \boldsymbol{\kappa}^2_k)$ and denote with $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$, $\mathbf{X} = (X_1, \ldots, X_T)$ and $\boldsymbol{s} = (s_1, \ldots, s_T)^\top$ the collection of response variables, covariates and state variables, respectively.

Since the joint posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$ is not tractable, we follow a data augmentation strategy and introduce the joint posterior $p(\boldsymbol{\theta}, \mathbf{s}|\mathbf{y}, \mathbf{X})$. We sample groups of parameters and latent variables from their full conditional distributions, following a block Gibbs scheme. Our sampling strategy deviates in three ways from the one in [32]. First, since we include the global shrinkage parameter τ not only in the prior for $\boldsymbol{\gamma}$ but also in the prior for $\boldsymbol{\beta}$, the full conditional distribution of τ depends on the tensor coefficients. Second, we integrate out $\boldsymbol{\gamma}$ from the full conditional of $\boldsymbol{\beta}$ to allow $\boldsymbol{\beta}$ to depend directly on the observed data. The resulting collapsed Gibbs sampler allows us to achieve exact sampling for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and to improve the sampler efficiency [33]. Third, we apply random scan Gibbs to increase the efficiency of the sampler [47].

At the first step of the Gibbs sampler, the $B_{\ell,k}$s and the PARAFAC margins are drawn from their full conditional distributions. The back-fitting sampling strategy allows sampling from tractable distributions, i.e. conditionally normal distributions, and for splitting the parameter vector into blocks.

At every iteration of the sampling algorithm, we randomly select a subset of component indices $\{d_1, \ldots, d_{D^*}\}$ of fixed size D^* from the set $\{1, 2, \ldots, D\}$, where $D^* < D$ and a subset of mode indices $\{m_1, \ldots, m_{M^*}\}$ of fixed size M^*, where $M^* < M$ from the set $\{1, 2, \ldots, M\}$. For $k \in \{1, \ldots, K\}$ and $\ell \in \{1, \ldots, N\}$, all the elements of $B_{\ell,k}$ and the PARAFAC margins $\gamma^{(d)}_{\ell,m,j_m,k}$ are sampled from their full conditional distributions:

1. Draw $\boldsymbol{\beta}^{(d)}_{\ell,m,j_m,k}$ from $f(\boldsymbol{\beta}^{(d)}_{\ell,m,j_m,k}|\mathbf{y}, \boldsymbol{X}, \mu_{\ell,k}, \boldsymbol{\beta}_{\ell,-j_m,k}, \sigma^2_{\ell,k}, \tau_{\ell,k}, \zeta^{(d)}_{\ell,k}, w^{(d)}_{\ell,m,j_m,k}, \kappa^2_{\ell,m,k})$ for $d \in \{d_1, \ldots, d_{D^*}\}$ and $m \in \{m_1, \ldots, m_{M^*}\}$ which is a multivariate normal distribution, where the $\{d_1, \ldots, d_{D^*}\}$ and $\{m_1, \ldots, m_{M^*}\}$ have been randomly selected according to Algorithm 1.

2. Draw $\gamma^{(d)}_{\ell,m,j_m,k}$ from $f(\gamma^{(d)}_{\ell,m,j_m,k}|\boldsymbol{\beta}^{(d)}_{m,j_m}, \tau_{\ell,k}, \zeta^{(d)}_{\ell,k}, w^{(d)}_{\ell,m,j_m,k}, \kappa^2_{\ell,m,k})$, which is a univariate normal distribution.

Let us denote the Inverse Gamma and the Generalized Inverse Gaussian distributions with IG and GIG, respectively. The Gibbs updates for the remaining parameters and hyper-parameters are:

3. Draw $\zeta^{(d)}_{\ell,m,k}$ from the GIG distribution $f(\zeta^{(d)}_{\ell,m,k}|\boldsymbol{\beta}^{(d)}_{\ell,k}, \boldsymbol{\gamma}^{(d)}_{\ell,k}, \boldsymbol{\kappa}^2_{\ell,k}, \boldsymbol{w}^{(d)}_{\ell,k})$.

4. Draw $\tau_{\ell,k}$ from the GIG distribution $f(\tau_{\ell,k}|\boldsymbol{\beta}_{\ell,k}, \boldsymbol{\gamma}_{\ell,k}, \boldsymbol{\kappa}^2_{\ell,k}, \boldsymbol{\zeta}_{\ell,k}, \boldsymbol{w}_{\ell,k})$.

5. Draw $\lambda^{(d)}_{\ell,m,k}$ from $f(\lambda^{(d)}_{\ell,m,k}|\boldsymbol{\gamma}^{(d)}_{\ell,m,k}, \tau_{\ell,k}, \zeta^{(d)}_{\ell,k})$ which is a Gamma distribution.

6. Draw $w^{(d)}_{\ell,m,j_m,k}$ from the GIG distribution $f(w^{(d)}_{\ell,m,j_m,k}|\gamma^{(d)}_{\ell,m,j_m,k}, \lambda^{(d)}_{\ell,m,k}, \tau_{\ell,k}, \zeta^{(d)}_{\ell,k})$.
7. Draw $\kappa^2_{\ell,m,k}$ from the GIG distribution $f(\kappa^2_{\ell,m,k}|\boldsymbol{\beta}_{\ell,m,k}, \boldsymbol{\gamma}_{\ell,m,k}, \tau_{\ell,k}, \boldsymbol{\zeta}_{\ell,k})$.
8. Draw $\sigma^2_{\ell,k}$ from the IG distribution $f(\sigma^2_{\ell,k}|\mathbf{y}, \boldsymbol{X}, \mu_{\ell,k}, \boldsymbol{\beta}_{\ell,k})$.
9. Draw $\mu_{\ell,k}$ from the Gaussian distribution $f(\mu_{\ell,k}|\mathbf{y}, \boldsymbol{X}, \boldsymbol{\beta}_{\ell,k}, \sigma^2_{\ell,k})$.
10. Draw transition probabilities $(p_{1,k}, \dots, p_{K,k})$ from the Dirichlet distribution $f(p_{1,k}, \dots, p_{K,k}|\boldsymbol{s})$.

Regarding the hidden states, we apply a Forward-Filtering Backward-Sampling (FFBS) strategy.

11. Compute iteratively the vector of smoothed probabilities $\xi_{t|T} = p(s_t|\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{X})$ by using the Hamilton filter recursions, and draw the state vector s_t from the multinomial distribution $\mathcal{M}(1, \xi_{t|T})$.

The derivation of the full conditional distributions for the parameters and the FFBS recursions can be found in Appendix B.

The current implementation is a variation of the usual Gibbs with a random scan. More concretely, consider that of interest is the posterior distribution $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^n$, but at each iteration, only a random subset of fixed size, say $n^* \leq n$, of the parameter vector, is updated. Moreover, every index set of size n^* has an equal chance of being selected. We describe in Algorithm 1 the steps of the sampler, which we call a Random Partial Scan Gibbs (RPSG).

Algorithm 1 The steps in a Random Partial Scan Gibbs

```
S1: Draw uniformly $J \subset \{1, \dots, n\}$ a random set of distinct indices of size $n^* \leq n$ so that each subset has an equal chance of being selected.
S2: If $J = (j_1, \dots, j_{n^*})$, update $\boldsymbol{\theta}_J = (\theta_{j_1}, \theta_{j_2}, \dots, \theta_{j_{n^*}})$ using a random scan and leave the other components of $\boldsymbol{\theta}$ unchanged.
```

The transition kernel of the RPSG satisfies the detailed balance condition, hence:

Remark 1. The chain generated by the RPSG sampler described in Algorithm 1 is an ergodic Markov chain with stationary distribution π.

To illustrate the performance of our proposed algorithm for tensor regression, we carried out an extensive simulation study for both simple and Markov Switching tensor regression for different specifications of the number of regimes and of the PARAFAC rank (see Appendix C in the Supplementary for further details). We study the proposed MCMC algorithm's efficiency by examining the MCMC chain empirical autocorrelation function (ACF) and the mean square error (MSE) of the true and sampled coefficient values. The regression model and the MCMC algorithm provide reasonably accurate coefficient estimates (posterior mean), and the regimes are successfully recovered by the maximum a posteriori estimates in different experimental settings.

The inferential performances are similar when $D \in \{3, 5, 7\}$. Additional simulations are carried out as robustness checks in two directions. First, the true coefficients are contaminated with white noise in such a way that the ranks are considered full for all different coefficients. The MCMC procedure can recover the patterns of the true coefficients reasonably well for all values of $D \in \{3, 5, 7\}$ (Fig. C.5 in Appendix C). Second, we evaluate the performances of different models with a different number of regimes, $K \in \{2, 3\}$, using Watanabe-Akaike Information Criterion (WAIC). We report the results in Table C.4. WAIC also confirms the simulation results that when K is fixed, varying D doesn't change the score too much. When D is fixed, a smaller K does help improve the model performance with a much lower WAIC score.

4. Empirical application

We test the validity of our tensor regression model using two real-world applications. In particular, we show through the applications that our tensor regression model: i) outperforms the competing estimation methods (ordinary least squares and linear LASSO) in terms of both in-sample and out-of-sample fitting; ii) captures the structural / regime changes in the data by taking advantage of the latent Markov-switching process.

4.1. Volatility index of the US market

In the first application, we study the relationship between the daily volatility index of the US market, also known as VIX and the crude oil ETF volatility index (OVX) with several other financial indicators. This is motivated by the fact that VIX has been recognized as the key benchmark index for measuring the market's expectations and sentiments, and predicting VIX is a crucial step for traders and investors when developing their trading strategies. In this regard, [16] studied the long-range dependence in the VIX data by including a vector of the average of the logarithm of VIX for the last $h \in \{1, 5, 10, 22, 66\}$ days (to mirror daily, weekly, bi-weekly, monthly and quarterly component) in a family of heterogeneous autoregressive (HAR) processes.

We follow similar strategy as [16] in forecasting VIX, but adapt it into a multiple-equation tensor regression framework, where we regress VIX on OVX (11) and vice versa (12) together with other covariates: the h day log-return for S&P 500, exchange rate (proxy by US dollar index), spot price of WTI crude oil for $h \in \{1, \ldots, 44\}$. To take advantage of the tensor structure, we construct the covariates for each response variable as a 4×44 matrix, which implies that the coefficient to be estimated is also a 4×44 matrix. The model specification for the two variables $y_{1,t} = \text{VIX}_t$ and $y_{2,t} = \text{OVX}_t$ is:

$$\begin{cases} \text{VIX}_t = \mu_1(s_t) + \left\langle B_1(s_t), \begin{pmatrix} \text{SP}_{t-1} & \ldots & \text{SP}_{t-h} & \ldots & \text{SP}_{t-44} \\ \text{ER}_{t-1} & \ldots & \text{ER}_{t-h} & \ldots & \text{ER}_{t-44} \\ \text{Oil}_{t-1} & \ldots & \text{Oil}_{t-h} & \ldots & \text{Oil}_{t-44} \\ \text{OVX}_{t-1} & \ldots & \text{OVX}_{t-h} & \ldots & \text{OVX}_{t-44} \end{pmatrix} \right\rangle + \sigma_1(s_t)\varepsilon_{1t}, & (11) \\ \text{OVX}_t = \mu_2(s_t) + \left\langle B_2(s_t), \begin{pmatrix} \text{SP}_{t-1} & \ldots & \text{SP}_{t-h} & \ldots & \text{SP}_{t-44} \\ \text{ER}_{t-1} & \ldots & \text{ER}_{t-h} & \ldots & \text{ER}_{t-44} \\ \text{Oil}_{t-1} & \ldots & \text{Oil}_{t-h} & \ldots & \text{Oil}_{t-44} \\ \text{VIX}_{t-1} & \ldots & \text{VIX}_{t-h} & \ldots & \text{VIX}_{t-44} \end{pmatrix} \right\rangle + \sigma_2(s_t)\varepsilon_{2t}. & (12) \end{cases}$$

To guide the selection of the model with the best in-sample fitting and out-of-sample forecasting among all realistic combinations of the number of regimes and the number of PARAFAC components, we compute the Watanabe-Akaike Information Criterion (WAIC, 39). WAIC is more appealing than AIC and DIC since it accounts for model prediction performances and is well suited for a Bayesian setup as it can be easily computed using the MCMC samples from the posterior [18]. The WAIC is defined as $\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}})$ where lppd denotes the logarithm of the pointwise predictive density and p_{WAIC} is the correction term for the effective number of parameters which adjusts for model complexity.

In Table 1, we report the WAIC for models with different values of K and D together with the Mean Square Error (MSE) and Mean Absolute Error (MAE) for the in- and out-of-sample prediction and different horizons of 1 and 5 days. Furthermore, we report the 95% credible intervals for the regime-specific intercept $\mu_{\ell,k}$, $\ell \in \{1, 2\}$ and $k \in \{1, 2, 3\}$. In the estimation the constraint $\mu_{2,1} < \mu_{2,2} < \cdots < \mu_{2,K}$ has been assumed in order to achieve identification. The main findings can be summarized as follows. Tensor models (MSTR(K, D) and TR(D)=MSTR$(1, D)$ in Table 1) consistently outperform the Least Squares (LS) and linear LASSO across almost all measures. Markov-switching Tensor Regressions MSTR(K, D) outperform simple Tensor Regressions TR(D). We show the in-sample fitting results of an MSTR model (MSTR$(2, 2)$) against LS and LASSO in Fig. 3. See also the comparison of in-sample fitting of models MSTR$(3, 2)$ and MSTR$(3, 3)$ in the supplement. The in-sample fitting of the LS and Linear LASSO regressions fails to capture the structural changes in the series of VIX and OVX. However, these structural changes are successfully captured by an MSTR, for which we assumed two possible regimes represent high and low volatility levels. Furthermore, the three-regime and three-component model MSTR$(3, 3)$ has the best in-sample performance. In contrast, the two-regime model MSTR$(2, 3)$ outperforms the best out-of-sample at the two horizons considered.

In addition, the regime separation is better supported by models with two regimes, MSTR$(2, 2)$ and MSTR$(2, 3)$, than in the three-regime models, MSTR$(3, 2)$ and MSTR$(3, 3)$. Note that the posterior credible intervals of the second equation intercept do not overlap across the two regimes (boldfaced intervals in Table 1). Between the two-regime models, we chose MSTR$(2, 2)$ for the data analysis because it is preferred by the *WAIC* criterion over MSTR$(2, 3)$. Fig. 3 shows that the MSTR$(2, 2)$ identified two regimes with distinctive regime-specific intercepts. Regime 2, representing a high level of Oil volatility, has a higher intercept value than Regime 1, which represents a low level of Oil volatility. The regime separation can be further described by inspecting the estimated effects of h-day log-return of oil prices and S&P 500 on VIX (blue dots) and OVX (red dots), respectively, in Fig. 4. The dots in the plots correspond to the values of parameters in the low-volatility ($s_t = 1$) and in the high-volatility ($s_t = 2$) regimes. The 90% HPD regions (grey ellipses) provide evidence of coefficient heterogeneity across regimes (asymmetric effects), equations (market asymmetry) and lags (long-term effects).

Regarding the asymmetric effects, we found evidence of the limited impact of the h-day oil and S&P 500 log-returns on both VIX and OVX in the low-volatility regime.

The values of coefficients are mostly centred around zero in this regime. As for the market asymmetry, the returns on oil have a stronger effect on OVX than on VIX in the high-volatility regime.

Table 1: Model Comparison for Financial Application. The table compares WAIC, MSE, MAE and credible intervals of the intercepts between models and regression methods with a different number of hidden regimes (K) and a different number of components (D). For MSE and MAE, we report the results for both in-sample and out-of-sample fitting in panel (a) (with forecasting horizons $h = 1$ and $h = 5$ days ahead). For the equation- and regime-specific intercepts $\mu_{\ell,k}$, we report their median $\hat{\mu}_{\ell,k}$ and their 2.5% and 97.5% quantiles $[\underline{\mu}_{\ell,k}, \bar{\mu}_{\ell,k}]$ in panel (b). *Note: MSTR(K, D) denotes the Markov-Switching Tensor Regression, TR(D) = $MSTR(1, D)$ the Tensor Regression, LS and LASSO the Least Squares and LASSO, respectively.*

(a) Predictive ability

Models	$WAIC$	In-sample		Out-of-sample			
				$h = 1$		$h = 5$	
		MSE	MAE	MSE	MAE	MSE	MAE
$TR(2)$	3231.06	0.3097	0.4324	0.2540	0.4232	0.3581	0.5182
MSTR(2, 2)	1907.28	0.0892	0.2376	0.1409	0.3342	0.1379	0.3063
$MSTR(3, 2)$	911.11	0.0339	0.1447	0.3199	0.4024	0.1976	0.3388
$TR(3)$	3282.48	0.3445	0.4534	0.2172	0.4155	0.2905	0.4751
MSTR(2, 3)	2347.39	0.1019	0.2505	**0.0939**	**0.2641**	**0.0659**	**0.2132**
MSTR(3, 3)	**518.13**	**0.0272**	**0.1221**	0.9328	0.9398	0.3471	0.5236
LS	–	0.3049	0.4266	0.1945	0.3474	0.3668	0.5211
$LASSO$	–	0.4207	0.5259	0.5199	0.6363	0.6940	0.7589

(b) Intercept estimates and credible intervals

Models	$(\hat{\mu}_{1,1}, \hat{\mu}_{2,1})$ $([\underline{\mu}_{1,1}, \bar{\mu}_{1,1}], [\underline{\mu}_{2,1}, \bar{\mu}_{2,1}])$	$(\hat{\mu}_{1,2}, \hat{\mu}_{2,2})$ $([\underline{\mu}_{1,2}, \bar{\mu}_{1,2}], [\underline{\mu}_{2,2}, \bar{\mu}_{2,2}])$	$(\hat{\mu}_{1,3}, \hat{\mu}_{2,3})$ $([\underline{\mu}_{1,3}, \bar{\mu}_{1,3}], [\underline{\mu}_{2,3}, \bar{\mu}_{2,3}])$
$TR(2)$	$(0.0016, -4.4 \times 10^{-05})$ $([-0.0452, 0.0478], [-0.0295, 0.0308])$	–	–
MSTR(2, 2)	$(-0.3100, 0.0086)$ $([-0.3304, -0.2900]$, **[-0.0438, 0.0595]**)	$(0.4157, 0.3938)$ $([0.2660, 0.5520]$, **[0.2855, 0.5268]**)	–
$MSTR(3, 2)$	$(-0.3550, -0.0225)$ $([-0.3945, 0.2138], [-0.0635, 0.0145])$	$(-0.1373, 0.0026)$ $([-0.3744, 0.2609], [-0.0267, 0.0327])$	$(0.0988, 0.0268)$ $([-0.3633, 0.3168], [-0.0056, 0.0797])$
$TR(3)$	$(0.0003, -0.0006)$ $([-0.0532, 0.0506], [-0.0292, 0.0307])$	–	–
MSTR(2, 3)	$(-0.2285, 0.0050)$ $([-0.2545, -0.2029]$, **[-0.0227, 0.0330]**)	$(-0.0952, 0.4036)$ $([-0.2331, 0.0261]$, **[0.1849 0.5653]**)	–
MSTR(3, 3)	$(-0.0125, -0.0132)$ $([-0.1344, 0.0920]$, **[-0.1179, -0.0655]**)	$(-0.0258, 0.0024)$ $([-0.2565, 0.0504]$, **[-0.0216, 0.0262]**)	$(-0.0478, 0.0185)$ $([-0.2560, 0.0552]$, **[-0.0050, 0.0642]**)

There is also strong evidence of non-negligible long-term effects of oil prices (dark red points) on oil volatility, which is aligned with previous findings [3, 15]. The returns on the S&P 500 have a very limited effect on VIX in the low-volatility regime. In contrast, the effects of returns at all lags are more substantial during high-volatility periods. The impact of S&P 500 on OVX follows a similar pattern; the coefficients with medium lags tend to have a larger effect than lower and higher lags. From the shape of the ellipses, we can tell that the coefficients are mostly uncorrelated across regimes, with fewer coefficients showing small positive or negative correlations. The coefficient posterior variance in the low-volatility regime is generally larger than in the high-volatility.

4.2. Oil prices on stock returns

For the second application, we extend our matrix-variate tensor regression model to a 3-mode tensor regression model by constructing the covariates X_t as a three-dimensional array for each observation. Therefore, the coefficients B_ℓ also form a three-dimensional array with the same size as the covariates. In this application, we contribute to the debate on the interdependence between financial and oil markets [see, e.g., 40, 41] and examine the impact of oil price volatility on the stock market returns (S&P 500) at an aggregate level and on

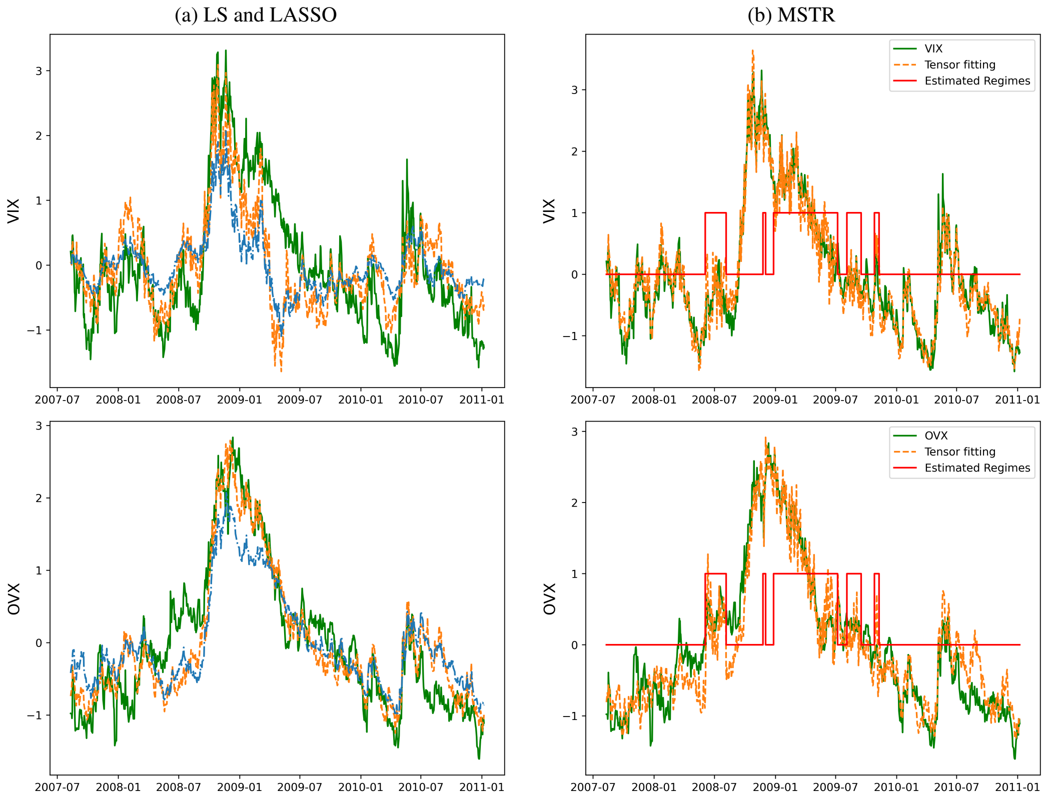

Fig. 3: Left: In-sample fitting for Least Squares (orange dashed) and LASSO (blue dashed). Right: In-sample fitting of the Markov-Switching Tensor Regression model MSTR(2, 2) (orange dashed) and estimated hidden states (red solid). The green solid line represents the VIX and VOX indexes (top and bottom).

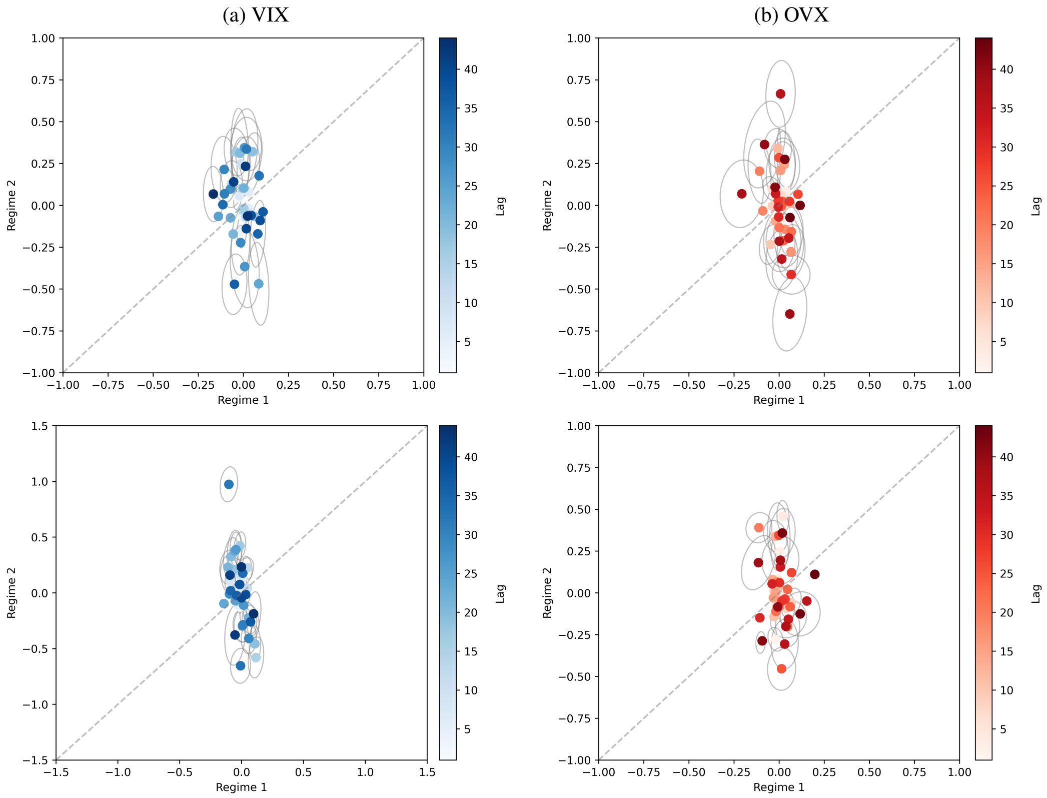

Fig. 4: Markov-switching Tensor Regression model MSTR(2, 2). Effects of h-day Oil (top) and S&P 500 (bottom) log-returns on VIX (left) and OVX (right) for $h \in \{1, \dots, 44\}$. Lighter and darker colors represent smaller and larger k, respectively. 90% Highest Posterior Density regions (gray ellipses) are plotted only for the coefficients which exhibit asymmetric effects across regimes (HPD ellipse does not intersect the 45° line).

the financial sector, energy sector and other sectors of S&P 500 at the disaggregate level. In particular, we classified the oil price volatility into Good Oil Volatility (GV), where the realized volatility is positive, and Bad Oil Volatility (BV), where the realized volatility is negative.

Our approach is different from the one in [40] in that we consider a Mixed Data Sampling (MIDAS) [34] framework by taking advantage of the tensor structure of the covariates. Our tensor regression setting naturally accommodates the multi-array structure of the covariates when data are sampled at different frequencies with different lags and reduces the number of parameters to be estimated by shrinking the unimportant parameters to small values. In particular, the response variable $y_{\ell,t}$ is the 4-week log-return of market ℓ at time t, where $\ell =$ 1 (S&P 500), 2 (financial sector), 3 (energy sector) and 4 (other S&P 500 sectors). The covariates are sampled weekly at the 1st week, 2nd week, 3rd week, 4th week before time t, indexed by $t-1/4, t-2/4, t-3/4, t-4/4$. Together with the GV and BV, the other covariates are the Exchange Rate Volatility (ER), TED Spread Volatility (IR) and VIX Index Volatility (VI), following a similar specification as in [40]. We arranged the different regressors along the rows (first mode) of the tensor covariates, weekly data points for four weeks along the columns (second mode), proceeding from the most to the least recent as usually done in the mixed-frequency literature. The weekly data points of the past months are stacked along the third axis (third mode). The MSTR model for this application is

$$y_{\ell,t} = \mu_\ell(s_t) + \sum_{j_3=1}^{4} \left\langle B_{\ell,\tilde{j}_3}(s_t), \begin{pmatrix} \mathrm{GV}^{(4)}_{t-\frac{1}{4}-j_3+1} & \mathrm{GV}^{(4)}_{t-\frac{1}{2}-j_3+1} & \mathrm{GV}^{(4)}_{t-\frac{3}{4}-j_3+1} & \mathrm{GV}^{(4)}_{t-j_3} \\ \mathrm{BV}^{(4)}_{t-\frac{1}{4}-j_3+1} & \mathrm{BV}^{(4)}_{t-\frac{1}{2}-j_3+1} & \mathrm{BV}^{(4)}_{t-\frac{3}{4}-j_3+1} & \mathrm{BV}^{(4)}_{t-j_3} \\ \mathrm{ER}^{(4)}_{t-\frac{1}{4}-j_3+1} & \mathrm{ER}^{(4)}_{t-\frac{1}{2}-j_3+1} & \mathrm{ER}^{(4)}_{t-\frac{3}{4}-j_3+1} & \mathrm{ER}^{(4)}_{t-j_3} \\ \mathrm{IR}^{(4)}_{t-\frac{1}{4}-j_3+1} & \mathrm{IR}^{(4)}_{t-\frac{1}{2}-j_3+1} & \mathrm{IR}^{(4)}_{t-\frac{3}{4}-j_3+1} & \mathrm{IR}^{(4)}_{t-j_3} \\ \mathrm{VI}^{(4)}_{t-\frac{1}{4}-j_3+1} & \mathrm{VI}^{(4)}_{t-\frac{1}{2}-j_3+1} & \mathrm{VI}^{(4)}_{t-\frac{3}{4}-j_3+1} & \mathrm{VI}^{(4)}_{t-j_3} \end{pmatrix} \right\rangle + \sigma_\ell(s_t)\varepsilon_{\ell,t}, \tag{13}$$

where $\tilde{j}_3 = \{(j_1, j_2, j_3), j_h \in \{1, \ldots, p_h\}, \forall h \neq 3\}$ and $B_{\ell,\tilde{j}_3}(s_t)$ denotes the j_3th slice of tensor coefficients $B_\ell(s_t)$ along the third mode. The conditional mean of the model in (13) is given as the sum over slices corresponding to different temporal lags (third mode).

Fig. D.4 of the Supplement shows the in-sample fitting of Least Squares and Linear LASSO (left column) and the in-sample fitting of MSTR (right column). Notably, Least Squares and linear LASSO fail to capture the volatility changes in the market return. In contrast, the MSTR can identify the most relevant episodes of market disruptions at both aggregate and disaggregate levels. For the aggregate analysis, when S&P 500 is used as the dependent variable, MSTR can identify the biggest disruption in the financial market in recent years, the 2008 global financial crisis.

For the disaggregated S&P 500 analysis (bottom plot and Fig. D.4 of the Supplement), when sector indices are used as dependent variables, MSTR can identify more episodes of market disruptions, including the 1997 Asia financial crisis, 2001 9/11 terrorist attack and 2002 corporate scandals and dot-com bubble together with the 2008 global financial crisis. The fact that MSTR can capture more structural changes at a disaggregated level can be largely attributed to the heterogeneity between different sectors. Thus, MSTR can also be an effective data integration tool.

Fig. 5 shows the effects of GV and BV on financial and energy sector log-returns (see also Fig. D.5 of the Supplement). We use different symbols to represent the weekly data sampled at different weeks for different lags $h = \{1, 2, 3, 4\}$, with ●: $t-(1+4(h-1))/4$, ✚: $t-2(1+4(h-1))/4$, ◆: $t-3(1+4(h-1))/4$ and ★ : $t-4(1+4(h-1))/4$. Lighter (darker) blue represents lower (higher) lag h. Coefficients with 90% HPD regions (grey ellipses) indicate large asymmetric effects.

For both aggregate and disaggregate analyses, GV and BV show more pronounced effects in the high-volatility regime ($s_t = 2$) than in the low-volatility regime ($s_t = 1$). This confirms the hypothesis of the financialization of the oil market [40, 41]. The HPD regions are more concentrated along the horizontal axis, most likely due to the smaller number of observations in regime 2 compared to regime 1. Regarding the long-term effects, GV has a more considerable impact on the markets at lower lags, while BV has a larger effect at higher lags. Similar asymmetries in the long and short-term impact have been documented within a univariate quantile regression framework by [41].

We report the MSE, MAE for the in-sample fitting and the out-of-sample forecasting with prediction horizons of 1-month and 5-month in the lower panel of Table D.1 of the Supplement. Overall, tensor regression offers competing performances with LS and LASSO, and MSTR performs strictly better in terms of in-sample fitting and short-term forecasting.

5. Conclusion

In this paper, we propose a new multiple-equation Markov-Switching Tensor Regression Model (MSTR) to work with high dimensional data where a common hidden Markov chain process introduces dependencies

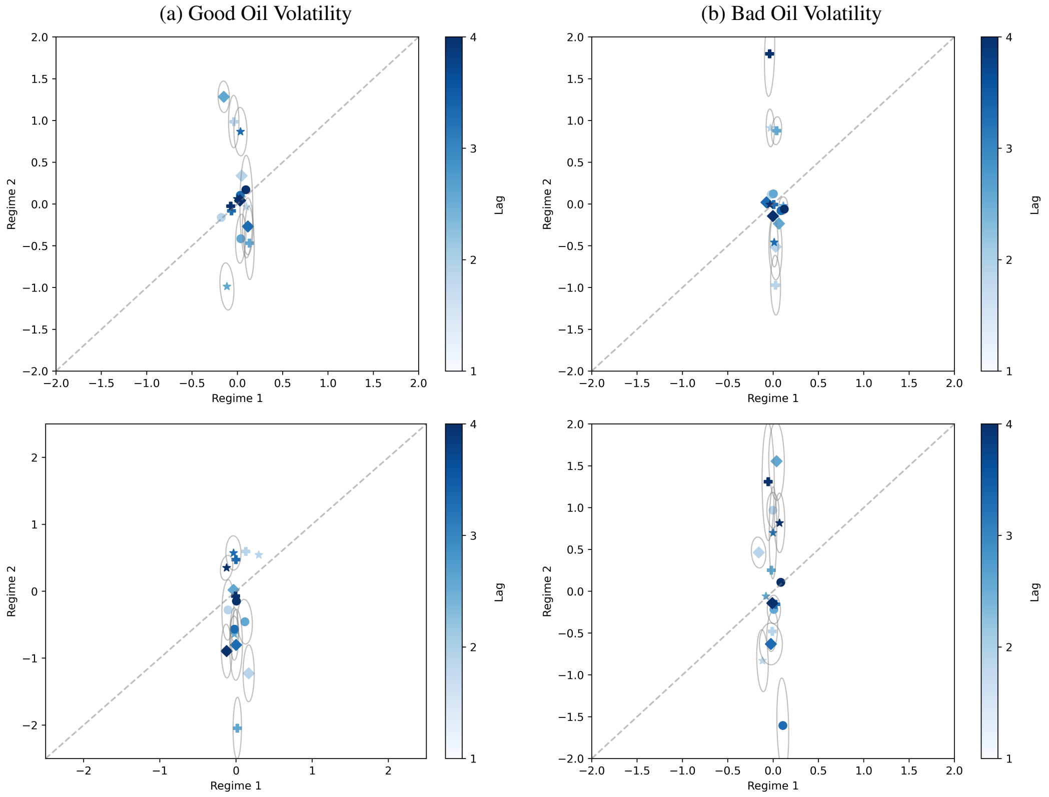

Fig. 5: The scatter plots show the effects of Good Oil Volatility (left) and Bad Oil Volatility (right) on the financial sector (top) and energy sector (bottom). Different symbols represent different weeks, with ●: $t-(1+4(h-1))/4$, ✚: $t-2(1+4(h-1))/4$, ◆: $t-3(1+4(h-1))/4$ and ★ : $t-4(1+4(h-1))/4$ for $h \in \{1,2,3,4\}$. Different blue shades represent lags, from order 1 (lighter) to order 4 (darker).

between equations and allows for latent regime changes and dynamic coefficients. A low-rank representation of the tensor coefficient is used to achieve dimensionality reduction. A hierarchical prior distribution is imposed to introduce further shrinkage effects in the regression model with many regressors. Multiple prior stages allow smoothing of the effects of the low-rank representation (soft PARAFAC decomposition). We developed an MCMC sampler based on Random Partial Scan Gibbs and a back-fitting strategy. We show that the Markov chain generated by the proposed sampler is stationary and converges to the target distribution. The validity and efficiency of the sampler are demonstrated using simulations with different settings. We also tested our MSTR with two real-world applications, where MSTR outperforms the competing algorithms in both in-sample fitting and out-of-sample forecasting. Moreover, MSTR provides more insight into the possible structural changes in the parameters by identifying regimes with regime-specific intercepts and variances, which are prevalent in time series data. The multiple-equation MSTR can also capture the heterogeneity in the data at aggregate and disaggregate levels to exploit more information in the data.

The proposed model and inference are ready to be used for tensor regression with covariate tensors of order 2 and 3. It can be considered in many other applications where regression on high-dimensional data is needed, and overparametrization or overfitting issues must be handled.

Acknowledgments and Funding

We thank the Editor, the Associate Editor and two anonymous referees for their comments and suggestions that have helped us improve the manuscript. Also, we would like to thank all the conference participants for helpful discussions at the 'International Society for Bayesian Analysis World Meeting', 2024, Venice, Italy, the 'Annual Meeting of the Statistical Society of Canada', 2024, St. John's, Newfoundland and Labrador. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. This research used the HPC-VERA multiprocessor cluster system at Ca' Foscari University of Venice. The first author was supported by the MUR - PRIN project '*Discrete random structures for Bayesian learning and prediction*' under g.a. n. 2022CLTYP4 and the Next Generation EU - '*GRINS - Growing Resilient, INclusive and Sustainable*' project (PE0000018), National Recovery and Resilience Plan (NRRP) - PE9 - Mission 4. The second author was supported by NSERC of Canada discovery grants RGPIN-2018-249547 and RGPIN-2024-04506.

A. Proofs

Before proving the main result in Proposition 1, let us recall some useful properties of conditionally independent normal random variables.

Remark A.1. Let $X|Z_1,R_1 \sim \mathcal{N}(Z_1,R_1)$ and $Y|Z_2,R_2 \sim \mathcal{N}(Z_2,R_2)$ be two conditionally independent normal random variables given Z_i,R_i, $1 \leq i \leq 2$, where $Z_i|Q_i \sim \mathcal{N}(0,Q_i)$, $i \in \{1,2\}$ are conditionally independent random variables given Q_1,Q_2, and R_i,Q_j, $i,j \in \{1,2\}$ are positive, possibly dependent, random variables. It can be shown that:

1. (marginal normal distribution) the marginal distributions are normal, i.e. $X|R_1,Q_1 \sim \mathcal{N}(0,R_1+Q_1)$ and $Y|R_2,Q_2 \sim \mathcal{N}(0,R_2+Q_2)$;

2. (marginal independence) the joint marginal $f_{XY|R_1,R_2,Q_1,Q_2}(x,y|r_1,r_2,q_1,q_2)$ can be written in the integral form $\int\int f_{X|Z_1,R_1}(x|z_1,r_1)f_{Y|Z_2,R_2}(y|z_2,r_2)\ f_{Z_1|Q_1}(z_1|q_1)\ f_{Z_2|Q_2}(z_2|q_2)dz_1dz_2$= $\int f_{X|Z_1,R_1}(x|z_1,r_1)f_{Z_1|Q_1}(z_1|q_1)dz_1\ \int f_{Y|Z_2,R_2}(y|z_2,r_2)f_{Z_2|Q_2}(z_2|q_2)dz_2$ and thus factorises as follows $f_{XY|R_1,R_2,Q_1,Q_2}(x,y|r_1,r_2,q_1,q_2) = f_{X|R_1,Q_1}(x|r_1,q_1)\ f_{Y|R_2,Q_2}(y|r_2,q_2)$, where $f_{X|R_1,Q_1}(x|r_1,q_1)$ and $f_{Y|R_2,Q_2}(y|r_2,q_2)$ are the densities of the two normal distributions given in 1.

From the two properties above, it follows that

3. By the law of iterated expectations and conditionally independence assumption E$[XY]$ = E$\{$E$[XY|Z_1,Z_2,R_1,R_2,Q_1,Q_2]\}$= E$\{E[X|Z_1,R_1]E[Y|Z_2,R_2]\}$= E$\{E[Z_1|Q_1]E[Z_2|Q_2]\} = 0$;

4. V(XY) = E$[X^2Y^2] - \{$E$[XY]\}^2$= E$\{$E$[X^2Y^2|Z_1,Z_2,R_1,R_2,Q_1,Q_2]\}$=E$\{E[X^2|Z_1,R_1,Q_1]E[Y^2|Z_2,R_2,Q_2]\}$= E$[(R_1+Z_1^2)(R_2+Z_2^2)]$ = E$\{$E$[R_1+Z_1^2|R_1,Q_1]$E$[R_2+Z_2^2|R_2,Q_2]\}$ which is equal to E$[$V$(X|R_1,Q_1)$V$(Y|R_2,Q_2)]$.

We drop the state subscript index k and the equation subscript index ℓ in the following. Let us define $\boldsymbol{\zeta} = (\zeta^{(1)},\ldots,\zeta^{(d)},\ldots,\zeta^{(D)})^\top$, $\boldsymbol{w} = (w_{1,1}^{(1)},\ldots,w_{m,j_m}^{(d)},\ldots,w_{M,p_M}^{(D)})^\top$, and $\boldsymbol{\kappa}^2 = (\kappa_1^2,\ldots,\kappa_M^2)^\top$. Assume for simplicity and w.l.o.g. that $M=2$. From (4) and (5) the $(j_1,\ldots,j_M)$th element of $B_1^{(d)}$ and $B_2^{(d)}$ are the two conditionally independent random variables $\beta_{1,\tilde{j}}^{(d)}|\boldsymbol{\gamma}^{(d)},\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w} \sim \mathcal{N}(\gamma_{1,j_1}^{(d)},\tau\kappa_1^2\zeta^{(d)})$ and $\beta_{2,\tilde{j}}^{(d)}|\boldsymbol{\gamma}^{(d)},\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w} \sim \mathcal{N}(\gamma_{2,j_2}^{(d)},\tau\kappa_2^2\zeta^{(d)})$, respectively. From the PARAFAC representation in (3), the $\tilde{j}$th element of $B^{(d)}$ can be written as the product of $\beta_{1,\tilde{j}}^{(d)}$ and $\beta_{2,\tilde{j}}^{(d)}$ with $\tilde{j} = (j_1,\ldots,j_M)$.

Remark A.2. From (5) and (6) and the assumption in (7), the results in Remark A.1 are applied with $X = \beta_{1,\tilde{j}}^{(d)}$, $Y = \beta_{2,\tilde{j}}^{(d)}$, $Z_1 = \gamma_{1,j_1}^{(d)}$, $Z_2 = \gamma_{2,j_2}^{(d)}$, $R_1 = \tau\kappa_1^2\zeta^{(d)}$, $R_2 = \tau\kappa_2^2\zeta^{(d)}$ $Q_1 = \tau\zeta^{(d)}w_{1,j_1}^{(d)}$ and $Q_2 = \tau\zeta^{(d)}w_{2,j_2}^{(d)}$ to obtain

1. $\beta_{1,\tilde{j}}^{(d)}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w} \sim \mathcal{N}(0,\tau\zeta^{(d)}(\kappa_1^2+w_{1,j_1}^{(d)}))$ and $\beta_{2,\tilde{j}}^{(d)}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w} \sim \mathcal{N}(0,\tau\zeta^{(d)}(\kappa_2^2+w_{2,j_2}^{(d)}))$ conditionally independent.

2. V$(\beta_{1,\tilde{j}}^{(d)}\beta_{2,\tilde{j}}^{(d)}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w})$ = V$(\beta_{1,\tilde{j}}^{(d)}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w})V(\beta_{2,\tilde{j}}^{(d)}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w})$ a.s. in τ, $\boldsymbol{\zeta}$, $\boldsymbol{\kappa}$, $\boldsymbol{w}$.

Proof of Proposition 1

We denote with $\beta_{m,\tilde{j}}^{(d)}$ the $\tilde{j}$th element of $B_m^{(d)}$, where $\tilde{j} = (j_1,\ldots,j_M)$ is a multiple-index. The variance of the coefficient entries of the soft PARAFAC can be written as:

$$\begin{aligned}
\mathrm{V}\left(B_{\tilde{j}}\right) &= \mathrm{E}\left\{\mathrm{V}\left(B_{\tilde{j}}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w}\right)\right\} = \mathrm{E}\left\{\mathrm{V}\left(\sum_{d=1}^{D}\prod_{m=1}^{M}\beta_{m,\tilde{j}}^{(d)}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w}\right)\right\} = \mathrm{E}\left\{\sum_{d=1}^{D}\prod_{m=1}^{M}\mathrm{V}\left(\beta_{m,\tilde{j}}^{(d)}|\tau,\boldsymbol{\zeta},\boldsymbol{\kappa},\boldsymbol{w}\right)\right\} \\
&= \mathrm{E}\left\{\sum_{d=1}^{D}\prod_{m=1}^{M}\tau\zeta^{(d)}\left(\kappa_m^2+w_{m,j_m}^{(d)}\right)\right\} = \mathrm{E}\left\{\tau^M\right\}\mathrm{E}\left\{\sum_{d=1}^{D}\left(\zeta^{(d)}\right)^M\right\}\mathrm{E}\left\{\prod_{m=1}^{M}\left(\kappa_m^2+w_{m,j_m}^{(d)}\right)\right\} \\
&= \frac{\Gamma(a_\tau+M)}{\Gamma(a_\tau)b_\tau^M}D\prod_{r=0}^{M-1}\frac{\alpha/D+r}{\alpha+r}\left(\frac{a_\kappa}{b_\kappa}+\frac{2b_\lambda^2}{(a_\lambda-1)(a_\lambda-2)}\right)^M. \qquad \text{(A.1)}
\end{aligned}$$

From the first line to the second line of (A.1) we used null mean and the conditional independence properties of $\beta_{m,\tilde{j}}^{(d)}$ across d and m from 1) in Remark A.2. It follows that $\prod_{m=1}^{M}\beta_{m,\tilde{j}}^{(d)}$ $d \in \{1,\ldots,D\}$ are conditionally independent across d and that the conditional variance of their sum is the sum of their conditional variances. Furthermore, from 2) in Remark A.2, the variance of the product of $\beta_{m,\tilde{j}}^{(d)}$ is equal to the product of their variances since they are conditionally independent and have a null mean. For the variance of the coefficient entries of hard PARAFAC, $\kappa_m^2 = 0$, thus

$$\mathrm{V}^{\text{hard}}\left(B_{\tilde{j}}\right) = \frac{\Gamma(a_\tau + M)}{\Gamma(a_\tau) b_\tau^M} D \prod_{r=0}^{M-1} \frac{\alpha/D + r}{\alpha + r} \left(\frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} \right)^M . \tag{A.2}$$

It is not hard to notice that a_κ/b_κ is the quantity that drives the additional variability of the soft PARAFAC by comparing (A.1) and (A.2). The goal is to set $\mathrm{V}(B_{\tilde{j}}) = V^*$ and $AV = AV^*$. By exploiting $\mathrm{V}(B_{\tilde{j}})/\mathrm{V}^{\text{hard}}(B_{\tilde{j}}) = (1 - AV^*)^{-1}$ we have

$$\frac{\mathrm{V}(B_{\tilde{j}})}{\mathrm{V}^{\text{hard}}(B_{\tilde{j}})} = \frac{\left(\frac{a_\kappa}{b_\kappa} + \frac{2b_\lambda^2}{(a_\lambda-1)(a_\lambda-2)} \right)^2}{\left(\frac{2b_\lambda^2}{(a_\lambda-1)(a_\lambda-2)} \right)^2} = \left(\frac{a_\kappa}{b_\kappa} \frac{(a_\lambda - 1)(a_\lambda - 2)}{2b_\lambda^2} + 1 \right)^2 = (1 - AV^*)^{-1} .$$

Solving the above equation for a_κ/b_κ and given a_κ/b_κ is positive we get

$$a_\kappa/b_\kappa = \left((1 - AV^*)^{-1/2} - 1 \right) 2b_\lambda^2/((a_\lambda - 1)(a_\lambda - 2)). \tag{A.3}$$

By setting $\mathrm{V}(B_{\tilde{j}}) = V^*$ we have,

$$2b_\lambda^2/((a_\lambda - 1)(a_\lambda - 2)) = b_\tau/a_\tau \sqrt{a_\tau V^*/((a_\tau + 1)C_\zeta)} - a_\kappa/b_\kappa. \tag{A.4}$$

Combing (A.3) and (A.4) we obtain $a_\kappa/b_\kappa = b_\tau/a_\tau \sqrt{a_\tau V^*/((a_\tau + 1)C_\zeta)} \left(1 - \sqrt{1 - AV^*} \right)$. □

Proof of Proposition 2

From the ℓth equation of the system (1) and the linearity property of the scalar product for tensors, see [21], it follows that

$$\begin{aligned}
\langle B_\ell(s_t), X_t \rangle &= \left\langle \sum_{d=1}^{D} B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t), X_t \right\rangle \\
&= \left\langle B_{\ell,m}^{(d)}(s_t), B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,m-1}^{(d)}(s_t) \circ B_{\ell,m+1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t) \circ X_t \right\rangle + \sum_{d' \neq d} \left\langle B_{\ell,1}^{(d')}(s_t) \circ \cdots \circ B_{\ell,M}^{(d')}(s_t), X_t \right\rangle \\
&= \sum_{j_m=1}^{p_m} \left\langle B_{\ell,m,\tilde{j}_m}^{(d)}(s_t), \left(B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,m-1}^{(d)}(s_t) \circ B_{\ell,m+1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t) \circ X_t \right)_{\tilde{j}_m} \right\rangle + R_{\ell,t}^{(d)}(s_t) \\
&= \boldsymbol{\beta}_{\ell,m,j_m}^{(d)}(s_t)^\top \mathrm{vec}(B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,m-1}^{(d)}(s_t) \circ B_{\ell,m+1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t) \circ X_t)_{\tilde{j}_m} + R_{\ell,m,j_m,t}^{(d)}(s_t) + R_{\ell,t}^{(d)}(s_t) \\
&= \boldsymbol{\beta}_{\ell,m,j_m}^{(d)}(s_t)^\top \Psi_{\ell,m,j_m,t}^{(d)}(s_t) + R_{\ell,m,j_m,t}^{(d)}(s_t) + R_{\ell,t}^{(d)}(s_t),
\end{aligned}$$

where

$$\begin{aligned}
R_{\ell,t}^{(d)}(s_t) &= \sum_{d' \neq d} \left\langle B_{\ell,1}^{(d')}(s_t) \circ \cdots \circ B_{\ell,M}^{(d')}(s_t), X_t \right\rangle, \\
R_{\ell,m,j_m,t}^{(d)}(s_t) &= \sum_{j'_m \neq j_m} \left\langle B_{\ell,m,\tilde{j}'_m}^{(d)}(s_t), \left(B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,m-1}^{(d)}(s_t) \circ B_{\ell,m+1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t) \circ X_t \right)_{\tilde{j}'_m} \right\rangle \\
&= \sum_{j'_m \neq j_m} \left\langle \left(B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t) \right)_{\tilde{j}'_m}, (X_t)_{\tilde{j}'_m} \right\rangle = \left\langle \left(B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t) \right)_{-j_m}, (X_t)_{-j_m} \right\rangle, \\
\Psi_{\ell,m,j_m,t}^{(d)}(s_t) &= \mathrm{vec}\left(B_{\ell,1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,m-1}^{(d)}(s_t) \circ B_{\ell,m+1}^{(d)}(s_t) \circ \cdots \circ B_{\ell,M}^{(d)}(s_t) \circ X_t \right)_{\tilde{j}_m} . \qquad \square
\end{aligned}$$

Proof of Remark 1

To simplify notation, we prove the result for $|I|$=3, but the result holds in general. Let $\mathcal{J} = \{1, \ldots, n\}$ be the set of parameter indices, suppose 3 distinct components with indices $j_1, j_2, j_3 \in \mathcal{J}$ are randomly selected to be updated, and denote with $\boldsymbol{j} = \{j_1, j_2, j_3\}$ the updated components and with $\boldsymbol{-j} = \mathcal{J} - \{j_1, j_2, j_3\}$ the components that are not updated. Furthermore, we assume the order in which the three components are updated is also random.

Given that the components in the complement of the set $\boldsymbol{j}$ remain unchanged, we have the transition function

$$\begin{aligned}\mathcal{K}\left(\theta^{(i+1)}|\theta^{(i)}\right) =&\xi_1\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i)}_{j_2},\theta^{(i)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i)}_{-\boldsymbol{j}}\right)\\
&+\xi_2\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i)}_{j_2},\theta^{(i)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\\
&+\xi_3\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i)}_{-\boldsymbol{j}}\right)\\
&+\xi_4\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_1},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\\
&+\xi_5\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i+1)}_{j_3},\theta^{(i)}_{j_2},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right)\\
&+\xi_6\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i+1)}_{j_3},\theta^{(i)}_{j_1},\theta^{(i)}_{-\boldsymbol{j}}\right)\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3},\theta^{(i)}_{-\boldsymbol{j}}\right),\end{aligned}$$

where ξ_i, $i\in\{1,\ldots,6\}$ are the probabilities of all possible orders in which d_j could be updated, and $\sum_{i=1}^{6}\xi_i=1$. To save space, we ignore $\theta^{(i)}_{-\boldsymbol{j}}$ from the equations for the rest of the proof. For the detailed balance condition, we must show $\mathcal{K}\left(\theta^{(i+1)}|\theta^{(i)}\right)\pi\left(\theta^{(i)}\right)=\mathcal{K}\left(\theta^{(i)}|\theta^{(i+1)}\right)\pi\left(\theta^{(i+1)}\right)$. The left side term can be expanded into:

$$\begin{aligned}&\mathcal{K}\left(\theta^{(i+1)}|\theta^{(i)}\right)\pi\left(\theta^{(i)}\right)=\Big\{\xi_1\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i)}_{j_2},\theta^{(i)}_{j_3}\right)\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_3}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2}\right)\\
&+\left(\xi_2\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i)}_{j_2},\theta^{(i)}_{j_3}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_2}\right)+\xi_5\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_2}\right)\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i+1)}_{j_3},\theta^{(i)}_{j_2}\right)\right)\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_3}\right)\\
&+\left(\xi_4\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_3}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_1}\right)+\xi_6\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_2}\right)\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i+1)}_{j_3},\theta^{(i)}_{j_1}\right)\right)\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3}\right)\\
&+\xi_3\pi\left(\theta^{(i+1)}_{j_2}|\theta^{(i)}_{j_1},\theta^{(i)}_{j_3}\right)\pi\left(\theta^{(i+1)}_{j_1}|\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_3}\right)\pi\left(\theta^{(i+1)}_{j_3}|\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2}\right)\Big\}\pi\left(\theta^{(i)}\right),\end{aligned}$$

where $\theta^{(i)}=\left(\theta^{(i)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i)}_{j_3}\right)$ and $\theta^{(i+1)}=\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3}\right)$. Equivalently

$$\begin{aligned}&\mathcal{K}\left(\theta^{(i+1)}|\theta^{(i)}\right)\pi\left(\theta^{(i)}\right)=\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3}\right)\Bigg\{\frac{\xi_1}{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2}\right)}\frac{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_2},\theta^{(i)}_{j_3}\right)}\frac{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_3}\right)}{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_3}\right)}\\
&+\frac{\xi_2}{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_3}\right)}\frac{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_2},\theta^{(i)}_{j_3}\right)}\frac{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i+1)}_{j_3}\right)}{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_2}\right)}+\frac{\xi_3}{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2}\right)}\frac{\pi\left(\theta^{(i)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_1},\theta^{(i)}_{j_3}\right)}\frac{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_3}\right)}{\pi\left(\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_3}\right)}\\
&+\frac{\xi_4}{\pi\left(\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3}\right)}\frac{\pi\left(\theta^{(i)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_1},\theta^{(i)}_{j_3}\right)}\frac{\pi\left(\theta^{(i)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_1},\theta^{(i+1)}_{j_2}\right)}+\frac{\xi_5}{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i+1)}_{j_3}\right)}\frac{\pi\left(\theta^{(i)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i+1)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_1},\theta^{(i)}_{j_2}\right)}\frac{\pi\left(\theta^{(i+1)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i+1)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_2},\theta^{(i+1)}_{j_3}\right)}\\
&+\frac{\xi_6}{\pi\left(\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3}\right)}\frac{\pi\left(\theta^{(i)}_{j_1},\theta^{(i)}_{j_2},\theta^{(i+1)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_1},\theta^{(i)}_{j_2}\right)}\frac{\pi\left(\theta^{(i)}_{j_1},\theta^{(i+1)}_{j_2},\theta^{(i+1)}_{j_3}\right)}{\pi\left(\theta^{(i)}_{j_1},\theta^{(i+1)}_{j_3}\right)}\Bigg\}\pi\left(\theta^{(i)}\right).\end{aligned}\tag{A.5}$$

The expression in (A.5) is symmetric in $\theta^{(i)}$ and $\theta^{(i+1)}$ when $\xi_1=\xi_2=\xi_3=\xi_4=\xi_5=\xi_6$. This proof can be easily generalized to an arbitrary randomly selected number of components as long as $\boldsymbol{j}\subset\{1,\ldots,n\}$. □

B. Full conditional derivations

B.1. Full conditional distribution of the hidden state variables

A multi-move sampling is applied to sample from the joint posterior distribution of the hidden state variables. We apply forward filtering and backward sampling [17]. Let us introduce the set of allocation variables $\boldsymbol{\xi}_t=(\xi_{1,t},\ldots,\xi_{K,t})$, with $\xi_{k,t}=\mathbb{I}(s_t=k)$. Using dynamic factorization, the full conditional distribution of the hidden state is

$$p(s_1,\ldots,s_T|\mathbf{y},\mathbf{X},\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\mu},\mathbf{p})\propto\prod_{t=1}^{T}p(\mathrm{y}_t|\mathbf{X}_t,B(s_t),\boldsymbol{\sigma}^2(s_t))\prod_{k=1}^{K}\prod_{l=1}^{K}p_{l,k}^{\xi_{k,t}\xi_{l,t-1}}.\tag{B.1}$$

B.2. Full conditional distribution of the transition probability

$$p\left((p_{1,i},\ldots,p_{K,i})|\mathbf{s}\right)\propto\prod_{t=1}^{T}\prod_{l=1}^{K}p_{l,i}^{\xi_{lt}\xi_{i,t-1}}\prod_{l=1}^{K}p_{l,i}^{\nu_l-1}\propto\prod_{l=1}^{K}p_{i,l}^{\sum_{t=1}^{T}\xi_{l,t}\xi_{i,t-1}+\nu_l-1},\tag{B.2}$$

which is proportional to Dirichlet distribution $\mathcal{D}ir(\bar{\nu}_1,\ldots,\bar{\nu}_K)$, where $\bar{\nu}_l=\nu_l+\sum_{t=1}^{T}\xi_{l,t}\xi_{i,t-1}$.

B.3. Full conditional distribution of the state-specific parameters

Given the conditional independence assumption, we drop the state subscript index k and equation subscript index ℓ for simplicity in the following. From the prior for $B_m^{(d)}$, we have $\boldsymbol{\beta}^{(d)}_{m,j_m} \sim \mathcal{N}_{q_m}(\gamma^{(d)}_{m,j_m}\boldsymbol{\iota}_{q_m}, \tau\kappa_m^2\zeta^{(d)}I_{q_m})$. The posterior of the unknowns of the model is given by

$$p(\boldsymbol{\beta}^{(d)}_{m,j_m}, \gamma^{(d)}_{m,j_m}, \sigma^2, \mu, \tau, \zeta^{(d)}, \lambda_m^{(d)}, \kappa_m^2, w^{(d)}_{m,j_m} \mid \mathbf{y}, \boldsymbol{X}). \tag{B.3}$$

We adopt an MCMC procedure based on Gibbs sampling to generate the unknowns from 3 blocks.

Block 1: Sampling $\zeta^{(d)}$ and τ from $p(\zeta^{(d)}, \tau \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{w})$

We first sample $\boldsymbol{\zeta}$ from the joint posterior by integrating out τ

$$\begin{aligned}
& p\left(\boldsymbol{\zeta} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{w}\right) \propto \pi\left(\boldsymbol{\zeta}\right) p\left(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\kappa}, \boldsymbol{w}, \zeta^{(d)}\right) = \pi\left(\boldsymbol{\zeta}\right) \int_{\mathbb{R}^+} p\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\kappa}, \tau\right) p\left(\boldsymbol{\gamma} \mid \boldsymbol{w}, \tau\right) \pi(\tau) d\tau \\
& = \prod_{d=1}^{D} \zeta^{(d)\,\frac{\alpha}{D}-1} \int_{\mathbb{R}^+} \left(\prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{j_m=1}^{p_m} \left(\tau\zeta^{(d)}\kappa_m^2\right)^{-\frac{q_m}{2}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\beta}^{(d)}_{m,j_m} - \gamma^{(d)}_{m,j_m}\boldsymbol{\iota}_{q_m}\right)^{\top} \left(\tau\zeta^{(d)}\kappa_m^2\right)^{-1} \left(\boldsymbol{\beta}^{(d)}_{m,j_m} - \gamma^{(d)}_{m,j_m}\boldsymbol{\iota}_{q_m}\right) \right\} \right. \\
& \left. \cdot \left(\tau\zeta^{(d)} w^{(d)}_{m,j_m}\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \frac{{\gamma^{(d)}_{m,j_m}}^{2}}{\tau\zeta^{(d)} w^{(d)}_{m,j_m}} \right\} \right) \tau^{a_\tau - 1} e^{-b_\tau \tau} d\tau \\
& \propto \prod_{d=1}^{D} \zeta^{(d)\,\frac{\alpha}{D}-1} \int_{\mathbb{R}^+} \left(\prod_{d=1}^{D} \left(\tau\zeta^{(d)}\right)^{-\frac{\sum_{m=1}^{M} p_m(q_m+1)}{2}} \exp\left\{ -\frac{1}{2\tau\zeta^{(d)}} C_d \right\} \right) \tau^{a_\tau - 1} e^{-b_\tau \tau} d\tau,
\end{aligned}$$

where we defined $C_d = \sum_{m=1}^{M} \sum_{j_m=1}^{p_m} \left(\|\boldsymbol{\beta}^{(d)}_{m,j_m} - \gamma^{(d)}_{m,j_m}\boldsymbol{\iota}_{q_m}\|^2/\kappa_m^2 + {\gamma^{(d)}_{m,j_m}}^{2}/w^{(d)}_{m,j_m} \right)$ and $\|\cdot\|$ denotes the Euclidean norm. Let $I_0 = \sum_{m=1}^{M} p_m(q_m+1) = M \prod_{m=1}^{M} p_m + \sum_{m=1}^{M} p_m$ then

$$p\left(\boldsymbol{\zeta} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{w}\right) \propto \prod_{d=1}^{D} \zeta^{(d)\,\frac{\alpha}{D} - \frac{I_0}{2} - 1} \int_{\mathbb{R}^+} \tau^{a_\tau - \frac{DI_0}{2} - 1} \exp\left\{ -b_\tau \tau - \frac{\sum_{d=1}^{D} C_d}{2\tau\zeta^{(d)}} \right\} d\tau.$$

By definition, $\sum_{d=1}^{D} \zeta^{(d)} = 1$ which yields $\sum_{d=1}^{D}(b_\tau\tau\zeta^{(d)}) = b_\tau\tau \sum_{d=1}^{D} \zeta^{(d)} = b_\tau\tau$, moreover, by letting $a_\tau = \alpha$ we obtain

$$p\left(\boldsymbol{\zeta} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{w}\right) \propto \int_{\mathbb{R}^+} \left(\prod_{d=1}^{D} \zeta^{(d)\,\frac{\alpha}{D} - \frac{I_0}{2} - 1} \right) \tau^{(\alpha - \frac{DI_0}{2}) - 1} \exp\left\{ -\frac{1}{2} \sum_{d=1}^{D} \left(\frac{C_d}{\tau\zeta^{(d)}} + 2b_\tau\tau\zeta^{(d)} \right) \right\} d\tau. \tag{B.4}$$

We recognize from (B.4) that the kernel of $\phi^{(d)} = \tau\zeta^{(d)}$ is a Generalized Inverse Gaussian distribution $\phi^{(d)} \sim \mathcal{GiG}\left(\alpha/D - I_0/2, 2b_\tau, C_d\right)$. We then obtain $\zeta^{(d)}$ by normalizing $\phi^{(d)}$ as follows: $\zeta^{(d)} = \phi^{(d)}/\sum_{d=1}^{D} \phi^{(d)}$ [see, e.g., 8].

The full conditional of τ can be derived as follows

$$\begin{aligned}
& p\left(\tau \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\kappa}, \boldsymbol{w}\right) \propto p\left(\boldsymbol{\beta} | \boldsymbol{\gamma}, \tau, \boldsymbol{\zeta}, \boldsymbol{\kappa}, \boldsymbol{w}\right) p\left(\boldsymbol{\gamma} | \tau, \zeta^{(d)}, \boldsymbol{w}\right) p(\tau) \\
& = \tau^{a_\tau - 1} e^{-b_\tau\tau} \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{j_m=1}^{p_m} \left(\tau\zeta^{(d)}\right)^{-\frac{q_m+1}{2}} \exp\left\{ -\frac{\|\boldsymbol{\beta}^{(d)}_{m,j_m} - \gamma^{(d)}_{m,j_m}\boldsymbol{\iota}_{q_m}\|^2}{2\tau\zeta^{(d)}\kappa_m^2} \right\} \left(w^{(d)}_{m,j_m}\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \frac{{\gamma^{(d)}_{m,j_m}}^{2}}{\tau\zeta^{(d)} w^{(d)}_{m,j_m}} \right\} \\
& \propto \tau^{a_\tau - \frac{DI_0}{2} - 1} e^{-b_\tau\tau} \prod_{d=1}^{D} \exp\left\{ -\frac{1}{2\tau\zeta^{(d)}} \sum_{m=1}^{M} \sum_{j_m=1}^{p_m} \frac{\|\boldsymbol{\beta}^{(d)}_{m,j_m} - \gamma^{(d)}_{m,j_m}\boldsymbol{\iota}_{q_m}\|^2}{\kappa_m^2} \right\} = \tau^{a_\tau - \frac{DI_0}{2} - 1} \exp\left\{ -\frac{1}{2} \left(\sum_{d=1}^{D} \frac{C_d}{\tau\zeta^{(d)}} + 2b_\tau\tau \right) \right\}.
\end{aligned}$$

Therefore, the full conditional of τ is also a Generalized Inverse Gaussian distribution

$$p\left(\tau \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\kappa}, \boldsymbol{w}\right) \propto \mathcal{GiG}\left(a_\tau - DI_0/2, 2b_\tau, \sum_{d=1}^{D} C_d/\zeta^{(d)} \right).$$

Block 2: Sampling $\lambda_m^{(d)}$ and $w^{(d)}_{m,j_m}$ from $p(\lambda_m^{(d)}, w^{(d)}_{m,j_m} | \gamma^{(d)}_{m,j_m}, \tau, \zeta^{(d)})$

Notice that by the construction of the prior distributions, $\gamma^{(d)}_{m,j_m}$ follows a double exponential distribution given $\lambda_m^{(d)}$, τ, $\zeta^{(d)}$, that is $\gamma^{(d)}_{m,j_m} \sim \mathcal{DE}\left(0, \sqrt{\tau\zeta^{(d)}}/\lambda_m^{(d)}\right)$. The full conditional of $\lambda_m^{(d)}$ can be written as

$$p\left(\lambda_m^{(d)} \mid \gamma^{(d)}_{m,j_m}, \tau, \zeta^{(d)}\right) \propto \pi(\lambda_m^{(d)}) p\left(\gamma^{(d)}_{m,j_m} \mid \lambda_m^{(d)}, \tau, \zeta^{(d)}\right) \propto \left(\tau\zeta^{(d)}\right)^{-\frac{p_m}{2}} \left(\lambda_m^{(d)}\right)^{a_\lambda + p_m - 1} \exp\left\{ -\left(\frac{\sum_{j_m=1}^{p_m} \left|\gamma^{(d)}_{m,j_m}\right|}{\sqrt{\tau\zeta^{(d)}}} + b_\lambda \right) \lambda_m^{(d)} \right\},$$

which is the kernel of the gamma distribution $\mathcal{G}a(a_\lambda + p_m, \sum_{j_m=1}^{p_m} |\gamma_{m,j_m}^{(d)}| / \sqrt{\tau\zeta^{(d)}} + b_\lambda)$. The full conditional of $w_{m,j_m}^{(d)}$ is proportional to

$$p\left(w_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)}\right) \propto \pi\left(w_{m,j_m}^{(d)}\right) p\left(\gamma_{m,j_m}^{(d)} \mid \lambda_m^{(d)}, \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right) \propto w_{m,j_m}^{(d)-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\lambda_m^{(d)2} w_{m,j_m}^{(d)} + \frac{\gamma_{m,j_m}^{(d)}{}^2}{\tau\zeta^{(d)} w_{m,j_m}^{(d)}}\right)\right\},$$

which is the kernel of the GIG distribution $\mathcal{G}i\mathcal{G}(1/2, \lambda_m^{(d)2}, \gamma_{m,j_m}^{(d)}{}^2/\tau\zeta^{(d)})$.

Block 3: Sampling $\boldsymbol{\beta}_{m,j_m}^{(d)}, \gamma_{m,j_m}^{(d)}, \mu, \sigma^2, \kappa_m^2$ *from* $p\left(\boldsymbol{\beta}_{m,j_m}^{(d)}, \gamma_{m,j_m}^{(d)}, \kappa_m^2, \mu, \sigma^2 \mid \mathbf{y}, \mathbf{X}\right)$

We derive the full conditional of $\boldsymbol{\beta}_{m,j_m}^{(d)}$ in a way such that it only depends on observed data by integrating out $\gamma_{m,j_m}^{(d)}$. The total number of $\boldsymbol{\beta}_{m,j_m}^{(d)}$ we need to sample is $D\sum_{m=1}^{M} p_m$ and their full conditional distribution $p\left(\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \mathbf{y}, \mathbf{X}, \mu, \boldsymbol{\beta}_{-j_m}, \sigma^2, \tau, \zeta^{(d)}, \kappa_m^2, w_{m,j_m}^{(d)}\right)$ is proportional to

$$\begin{aligned} & p\left(\mathbf{y} \mid \mathbf{X}, \mu, \boldsymbol{\beta}, \sigma^2\right) \int_{\mathbb{R}} p\left(\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}, \tau, \kappa_m^2, \zeta^{(d)}\right) p\left(\gamma_{m,j_m}^{(d)} \mid \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right) d\gamma_{m,j_m}^{(d)} \\ & \propto \prod_{t \in \mathcal{T}} \exp\left\{-\frac{1}{2}\frac{(y_t - \mu - \langle B, X_t \rangle)^2}{\sigma^2}\right\} \int_{\mathbb{R}} p\left(\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}, \tau, \zeta^{(d)}, \kappa_m^2\right) p\left(\gamma_{m,j_m}^{(d)} \mid \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right) d\gamma_{m,j_m}^{(d)}, \end{aligned}$$

where $\mathcal{T} \subset \{1, \ldots, T\}$ contains all the indexes of the observations such that $s_t = k$, $k \in \{1, \ldots, K\}$. Thanks to the result in Proposition 2, and defining $R_t^{(d)} = \sum_{d' \neq d} \left\langle B_1^{(d')} \circ \cdots \circ B_M^{(d')}, X_t \right\rangle$, $R_{\ell,m,j_m,t}^{(d)} = \left\langle (B_1^{(d)} \circ \cdots \circ B_M^{(d)})_{-j_m}, (X_t)_{-j_m} \right\rangle$, $\Psi_{m,j_m,t}^{(d)} = \operatorname{vec}\left((B_1^{(d)} \circ \cdots \circ B_{m-1}^{(d)} \circ B_{m+1}^{(d)} \circ \cdots \circ B_M^{(d)} \circ X_t)_{\tilde{j}_m}\right)$ and $\tilde{y}_t = y_t - \mu - R_{m,j_m,t}^{(d)} - R_t^{(d)}$, the terms at the exponent in the likelihood become:

$$\left(y_t - \mu - \boldsymbol{\beta}_{m,j_m}^{(d)}{}^\top \Psi_{m,j_m,t}^{(d)} - R_{m,j_m,t}^{(d)} - R_t^{(d)}\right)^2 = \tilde{y}_t^2 + ||\Psi_{m,j_m,t}^{(d)\prime} \boldsymbol{\beta}_{m,j_m}^{(d)}||^2 - 2\boldsymbol{\beta}_{m,j_m}^{(d)}{}^\top \Psi_{m,j_m,t}^{(d)} \tilde{y}_t,$$

and the likelihood can be written as

$$p\left(\mathbf{y} \mid \mathbf{X}, \mu, \boldsymbol{\beta}, \sigma^2\right) \propto \exp\left\{-\frac{1}{2\sigma^2} \sum_{t \in \mathcal{T}} \left(||\Psi_{m,j_m,t}^{(d)\top} \boldsymbol{\beta}_{m,j_m}^{(d)}||^2 - 2\boldsymbol{\beta}_{m,j_m}^{(d)}{}^\top \Psi_{m,j_m,t}^{(d)} \tilde{y}_t\right)\right\}.$$

For the integration part, given $p(\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}, \tau, \zeta^{(d)}, \kappa_m^2)$ and $p\left(\gamma_{m,j_m}^{(d)} \mid \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right)$ are normal then from Remark A.2 the marginal distribution is normal with mean $\mathrm{E}\left[\boldsymbol{\beta}_{m,j_m}^{(d)}\right] = \mathrm{E}\left\{\mathrm{E}\left[\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}\right]\right\} = \mathbf{0}$, and variance $\mathrm{V}(\boldsymbol{\beta}_{m,j_m}^{(d)}) = \mathrm{V}\left(\mathrm{E}\left[\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}\right]\right) + \mathrm{E}\left[\mathrm{V}\left(\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}\right)\right] = (\tau\zeta^{(d)} w_{m,j_m}^{(d)} + \tau\zeta^{(d)}\kappa_m^2) I_{q_m}$. Let $\xi = \tau\zeta^{(d)}\left(w_{m,j_m}^{(d)} + \kappa_m^2\right)$, then the full conditional of $\boldsymbol{\beta}_{m,j_m}^{(d)}$ can be written as follows

$$\begin{aligned} p\left(\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \mathbf{y}, \mathbf{X}, \mu, \sigma^2, \kappa_m^2, \zeta^{(d)}, \tau, w_{m,j_m}^{(d)}\right) & \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_{m,j_m}^{(d)}{}^\top \left(\sum_{t \in \mathcal{T}} \frac{\Upsilon_{m,j_m,t}^{(d)}}{\sigma^2} + \frac{1}{\xi} I_{q_m}\right) \boldsymbol{\beta}_{m,j_m}^{(d)} - 2\boldsymbol{\beta}_{m,j_m}^{(d)}{}^\top \sum_{t \in \mathcal{T}} \frac{\Psi_{m,j_m,t}^{(d)} \tilde{y}_t}{\sigma^2}\right]\right\} \\ & \propto \mathcal{N}\left(\left(\sum_{t \in \mathcal{T}} \frac{\Upsilon_{m,j_m,t}^{(d)}}{\sigma^2} + \frac{1}{\xi} I_{q_m}\right)^{-1} \sum_{t \in \mathcal{T}} \frac{\Psi_{m,j_m,t}^{(d)} \tilde{y}_t}{\sigma^2}, \left(\sum_{t \in \mathcal{T}} \frac{\Upsilon_{m,j_m,t}^{(d)}}{\sigma^2} + \frac{1}{\xi} I_{q_m}\right)^{-1}\right), \end{aligned}$$

where we defined $\Upsilon_{m,j_m,t}^{(d)} = \Psi_{m,j_m,t}^{(d)} \otimes \Psi_{m,j_m,t}^{(d)}$. The full conditional of $\gamma_{m,j_m}^{(d)}$ given $\boldsymbol{\beta}_{m,j_m}^{(d)}$ can be written as

$$\begin{aligned} & p\left(\gamma_{m,j_m}^{(d)} \mid \boldsymbol{\beta}_{m,j_m}^{(d)}, \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}, \kappa_m^2\right) \propto p\left(\boldsymbol{\beta}_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)}, \tau, \zeta^{(d)}, \kappa_m^2\right) p\left(\gamma_{m,j_m}^{(d)} \mid \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right) \\ & \propto \exp\left\{-\frac{1}{2\tau\zeta^{(d)}}\left[\frac{q_m w_{m,j_m}^{(d)} + \kappa_m^2}{w_{m,j_m}^{(d)} \kappa_m^2}\left(\gamma_{m,j_m}^{(d)} - \frac{w_{m,j_m}^{(d)}}{q_m w_{m,j_m}^{(d)} + \kappa_m^2} \boldsymbol{\beta}_{m,j_m}^{(d)}{}^\top \boldsymbol{\iota}_{q_m}\right)^2\right]\right\} \propto \mathcal{N}\left(\frac{w_{m,j_m}^{(d)}}{q_m w_{m,j_m}^{(d)} + \kappa_m^2} \boldsymbol{\beta}_{m,j_m}^{(d)}{}^\top \boldsymbol{\iota}_{q_m}, \frac{\tau\zeta^{(d)} w_{m,j_m}^{(d)} \kappa_m^2}{q_m w_{m,j_m}^{(d)} + \kappa_m^2}\right). \end{aligned}$$

The full conditional of κ_m^2 can be written as

$$\begin{aligned} p\left(\kappa_m^2 \mid \boldsymbol{\beta}_{m,j_m}^{(d)}, \boldsymbol{\gamma}_m^{(d)}, \tau, \boldsymbol{\zeta}\right) & \propto \prod_{d=1}^{D} \prod_{j_m=1}^{p_m} p\left(\boldsymbol{\beta}_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}, \tau, \zeta^{(d)}, \kappa_m^2\right) p\left(\kappa_m^2\right) \\ & = \left(\kappa_m^2\right)^{a_\kappa - D\frac{p_m q_m}{2} - 1} \exp\left\{-\frac{1}{2}\left(\frac{\sum_{d=1}^{D} \sum_{j_m=1}^{p_m} ||\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\iota}_{q_m}||^2}{\tau\zeta^{(d)} \kappa_m^2} + 2b_\kappa \kappa_m^2\right)\right\} \\ & \propto \mathcal{G}i\mathcal{G}\left(a_\kappa - Dp_m q_m/2, 2b_\kappa, \sum_{d=1}^{D} \sum_{j_m=1}^{p_m} ||\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\iota}_{q_m}||^2 / \tau\zeta^{(d)}\right). \end{aligned}$$

The full conditional of σ^2 can be written as:

$$p\left(\sigma^2 \mid \mathbf{y}, \boldsymbol{X}, \mu, \boldsymbol{\beta}\right) \propto p\left(\mathbf{y} \mid \boldsymbol{X}, \mu, \boldsymbol{\beta}, \sigma^2\right) p\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-(a_\sigma + \frac{T}{2})-1} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{2}\sum_{t=1}^{T} (y_t - \langle B, X_t\rangle - \mu)^2 + b_\sigma\right)\right\},$$

which is the kernel of the IG distribution $\mathcal{IG}(a_\sigma^*, b_\sigma^*)$, where $a_\sigma^* = a_\sigma + \frac{T}{2}$ and $b_\sigma^* = \frac{1}{2}\sum_{t=1}^{T}(y_t - \langle B, X_t\rangle - \mu)^2 + b_\sigma$. Finally, let $\mu^* = \sum_{t=1}^{T}(y_t - \langle B, X_t\rangle)\,\sigma_\mu^{*2}$ and $\sigma_\mu^{*2} = \left(T/\sigma^2 + 1/\sigma_\mu^2\right)^{-1}$, the full conditional of μ is:

$$\begin{aligned} p\left(\mu \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right) &\propto p\left(\mathbf{y} \mid \boldsymbol{X}, \mu, \boldsymbol{\beta}, \sigma^2\right)\pi(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[T\mu^2 - 2\mu\sum_{t=1}^{T}(y_t - \langle B, X_t\rangle)\right] - \frac{1}{2}\frac{\mu^2}{\sigma_\mu^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\left(\frac{T}{\sigma^2} + \frac{1}{\sigma_\mu^2}\right)\mu^2 - 2\mu\frac{\sum_{t=1}^{T}(y_t - \langle B, X_t\rangle)}{\sigma^2}\right]\right\} \propto \mathcal{N}\left(\mu^*, \sigma_\mu^{*2}\right). \end{aligned}$$

References

[1] K. M. Agudze, M. Billio, R. Casarin, F. Ravazzolo, Markov switching panel with endogenous synchronization effects, Journal of Econometrics 230 (2022) 281–298.

[2] A. Armagan, D. B. Dunson, J. Lee, Generalized double Pareto shrinkage, Statistica Sinica 23 (2013) 119–143.

[3] F. M. Bandi, B. Perron, Long memory and the relation between implied and realized volatility, Journal of Financial Econometrics 4 (2006) 636–670.

[4] L. Bauwens, J.-F. Carpantier, A. Dufays, Autoregressive moving average infinite hidden Markov-switching models, Journal of Business & Economic Statistics 35 (2017) 162–182.

[5] D. Bianchi, M. Billio, R. Casarin, M. Guidolin, Modeling systemic risk with Markov switching graphical SUR models, Journal of Econometrics 210 (2019) 58 – 74.

[6] M. Billio, R. Casarin, Beta autoregressive transition Markov-switching models for business cycle analysis, Studies in Nonlinear Dynamics & Econometrics 15 (2011).

[7] M. Billio, R. Casarin, M. Iacopini, Bayesian Markov-switching tensor regression for time-varying networks, Journal of the American Statistical Association 119 (2024) 109–121.

[8] M. Billio, R. Casarin, M. Iacopini, S. Kaufmann, Bayesian dynamic tensor regression, Journal of Business & Economic Statistics 41 (2023) 429–439.

[9] M. Billio, R. Casarin, F. Ravazzolo, H. Van Dijk, Interactions between Eurozone and US Booms and Busts: A Bayesian Panel Markov-switching VAR model, Journal of Applied Econometrics 31 (2016) 1352–1370.

[10] M. Billio, A. Monfort, C. P. Robert, Bayesian estimation of switching ARMA models, Journal of Econometrics 93 (1999) 229–255.

[11] B. S. Caffo, C. M. Crainiceanu, G. Verduzco, S. Joel, S. H. Mostofsky, S. S. Bassett, J. J. Pekar, Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk, NeuroImage 51 (2010) 1140–1149.

[12] R. Casarin, C. Foroni, M. Marcellino, F. Ravazzolo, Economic uncertainty through the lenses of a mixed-frequency Bayesian panel Markov switching model, Annals of Applied Statistics 12 (2018) 2559 – 2568.

[13] R. Casarin, D. Sartore, M. Tronzano, A Bayesian Markov-switching correlation model for contagion analysis on exchange rate markets, Journal of Business & Economic Statistics 36 (2018) 101–114.

[14] S. Chib, Calculating posterior distributions and modal estimates in Markov mixture models, Journal of Econometrics 75 (1996) 79–97.

[15] F. Corsi, A simple approximate long-memory model of realized volatility, Journal of Financial Econometrics 7 (2009) 174–196.

[16] M. Fernandes, M. C. Medeiros, M. Scharth, Modeling and predicting the cboe market volatility index, Journal of Banking & Finance 40 (2014) 1–10.

[17] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models, Springer, New York, 2006.

[18] A. Gelman, J. Hwang, A. Vehtari, Understanding predictive information criteria for Bayesian models, Statistics and Computing 24 (2014) 997–1016.

[19] S. Guha, A. Rodriguez, Bayesian regression with undirected network predictors with an application to brain connectome data, Journal of the American Statistical Association 116 (2021) 581–593.

[20] R. Guhaniyogi, S. Qamar, D. B. Dunson, Bayesian tensor regression, Journal of Machine Learning Research 18 (2017) 2733–2763.

[21] W. Hackbusch, Tensor Spaces and Numerical Tensor Calculus, Springer Series in Computational Mathematics, Springer International Publishing, 2019.

[22] J. D. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica (1989) 357–384.

[23] T. Hastie, R. Tibshirani, Bayesian backfitting, Statistical Science 15 (2000) 196–213.

[24] N. Hauzenberger, Flexible mixture priors for large time-varying parameter models, Econometrics and Statistics 20 (2021) 87–108.

[25] S. Kaufmann, K-state Switching models with time-varying transition distributions: Does loan growth signal stronger effects of variables on inflation?, Journal of Econometrics 187 (2015) 82–94.

[26] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM Review 51 (2009) 455–500.

[27] J. Kossaifi, Z. C. Lipton, A. Kolbeinsson, A. Khanna, T. Furlanello, A. Anandkumar, Tensor regression networks, The Journal of Machine Learning Research 21 (2020) 4862–4882.

[28] D. R. Kowal, D. S. Matteson, D. Ruppert, Dynamic shrinkage processes, Journal of the Royal Statistical Society Series B 81 (2019) 781–804.

[29] R. A. Levine, G. Casella, Optimizing random scan Gibbs samplers, Journal of Multivariate Analysis 97 (2006) 2071–2100.

[30] K. Mardia, J. Kent, J. Bibby, Multivariate Analysis, AcademicPress, New York (1979).

[31] M. Ohlson, M. R. Ahmad, D. Von Rosen, The multilinear normal distribution: Introduction and some basic properties, Journal of Multivariate Analysis 113 (2013) 37–47.

[32] G. Papadogeorgou, Z. Zhang, D. B. Dunson, Soft tensor regression., Journal of Machine Learning Research 22 (2021) 219–1.

[33] C. Robert, The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, volume 2 of Springer Texts in Statistics, Springer New York, 2007.

[34] A. Rodriguez, G. Puggioni, Mixed frequency models: Bayesian approaches to estimation and prediction, International Journal of Forecasting 26 (2010) 293–311.

[35] C. A. Sims, D. F. Waggoner, T. Zha, Methods for inference in large multiple-equation Markov-switching models, Journal of Econometrics 146 (2008) 255–274.

[36] M. E. P. So, K. Lam, W. K. Li, A stochastic volatility model with Markov switching, Journal of Business & Economic Statistics 16 (1998) 244–253.

[37] D. Spencer, R. Guhaniyogi, R. Shinohara, R. Prado, Bayesian tensor regression using the Tucker decomposition for sparse spatial modeling, arXiv preprint arXiv:2203.04733 (2022).

[38] K. Wang, Y. Xu, Bayesian tensor-on-tensor regression with efficient computation, Statistics and Its Interface 17 (2024) 199.

[39] S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, The Journal of Machine Learning Research 11 (2010).

[40] J. Xiao, Y. Wang, Good oil volatility, bad oil volatility, and stock return predictability, International Review of Economics & Finance 80 (2022) 953–966.

[41] J. Xiao, Y. Wang, D. Wen, The predictive effect of risk aversion on oil returns under different market conditions, Energy Economics 126 (2023) 106969.

[42] R. Yu, G. Li, Y. Liu, Tensor regression meets Gaussian processes, in: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 482–490.

[43] R. Yu, Y. Liu, Learning from multiway data: Simple and efficient tensor regression, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, PMLR, New York, New York, USA, 2016, pp. 373–381.

[44] Z. Zhang, G. I. Allen, H. Zhu, D. Dunson, Tensor network factorizations: Relationships between brain structural connectomes and traits, Neuroimage 197 (2019) 330–343.

[45] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, Tensor-variate Gaussian processes regression and its application to video surveillance, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1265–1269.

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

[47] K. Łatuszyński, G. O. Roberts, J. S. Rosenthal, Adaptive Gibbs samplers and related MCMC methods, The Annals of Applied Probability 23 (2013) 66 – 98.