

Markov Switching Multiple-equation Tensor Regressions

Roberto Casarin[†] Radu Craiu[‡] and Qing Wang[†]

[†]Ca' Foscari University of Venice, Italy

[‡]University of Toronto, Canada

July 2, 2024

Abstract

We propose a new flexible tensor model for multiple-equation regression that accounts for latent regime changes. The model allows for dynamic coefficients and multi-dimensional covariates that vary across equations. We assume the coefficients are driven by a common hidden Markov process that addresses structural breaks to enhance the model flexibility and preserve parsimony. We introduce a new Soft PARAFAC hierarchical prior 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. We developed theoretical results to help hyperparameter choice. An efficient MCMC algorithm based on random scan Gibbs and back-fitting strategy is developed to achieve better 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 when compared to the current benchmark, Lasso regression.

Keywords: Bayesian inference, Dimensionality Reduction, Markov Switching, MCMC, Tensor Regression

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 (Spencer et al., 2022; Guha and Rodriguez, 2021), biostatistics, financial networks (Billio et al., 2024), or even more generally, in time series (Billio et al., 2023). People are often interested in characterizing the relationship between a tensor predictor and a scalar outcome (Guha and Rodriguez, 2021) or tensor outcome (Wang and Xu, 2024). Tensor regression has been studied extensively in this regard 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 follow the parsimonious strategy for dynamic behaviour that is provided by Hidden Markov Models (HMM) (Frühwirth-Schnatter, 2006) also known as Markov Switching (MS) (Kaufmann, 2015; Billio et al., 2016; Casarin et al., 2018a,b; Bianchi et al., 2019; Agudze et al., 2022). Efforts have been made to address abrupt structural changes in temporal networks by Billio et al. (2024). 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 multi-fold. First, we extend the soft tensor linear regression of Guha and Rodriguez (2021); Papadogeorgou et al. (2021); Wang and Xu (2024) to an HMM or MS framework to accommodate structural breaks. 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 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-parametrization and overfitting issues. In the tensor regression framework, a simple and straightforward approach 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, Zhang et al. (2019) 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, Caffo et al. (2010) 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, Yu and Liu (2016) applied Tucker decomposition on the tensor coefficients and proposed a fast algorithm (Tensor Projected Gradient) to minimize the empirical loss function. Zhou et al. (2013) used PARAFAC decomposition, a special case of Tucker decomposition, on the tensor coefficients and relied on maximum likelihood estimation to perform neuroimaging data analysis. Kossaifi et al. (2020) used a neural network combined with Tucker representation to address multi-way data analysis.

Within the Bayesian paradigm, Yu et al. (2018); Zhao et al. (2014) proposed non-parametric methods based on Gaussian Process priors. In a scalar on tensor regression framework, Guhaniyogi et al. (2017) proposed a novel multi-way shrinkage prior on the Parallel Factor (PARAFAC) representation, also called CP or Poyadic decomposition (see Kolda and Bader (2009)) of the coefficient tensor. Their work was extended by Spencer et al. (2022), who explored a more general Tucker decomposition on the tensor coefficients. In follow-up work, Guha and Rodriguez (2021) proposed a Bayesian network shrinkage prior and used a spike-and-slab prior to determine which brain nodes are most influential to creativity. In the case of tensor on tensor regression, Wang and Xu (2024) 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 (Papadogeorgou et al., 2021) where the hierarchical prior distribution of Guhaniyogi et al. (2017) 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 (Papadogeorgou et al., 2021) and Guhaniyogi et al. (2017) to improve the tractability of the conditional posterior distributions and developed an efficient MCMC algorithm to achieve better scalability. The latter relies on a random scan Gibbs sampler within the back-fitting strategy, usually employed in Bayesian tensor regression models (Levine and Casella, 2006; Łatuszyński et al., 2013; Hastie and Tibshirani, 2000).

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 (MSMETR) 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 matrix-valued and 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. In our model, we assume a system of N equations has time-varying parameters

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

$t = 1, 2, \dots, T$ where $y_{\ell t}$ is a scalar response variable, X_t is a $n \times m$ matrix of covariates, $B_\ell(s_t)$ are $n \times m$ coefficient matrices, $\varepsilon_{\ell t}$ are i.i.d from $\mathcal{N}(0, 1)$, $\{s_t\}_{t=1}^T$ is a common latent process, and $\langle \cdot, \cdot \rangle$ denotes the inner product for tensors (Kolda and Bader (2009)).

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

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

Alternative parameterizations for the coefficients can be used; see Frühwirth-Schnatter (2006) for conditionally linear single-equation models and Casarin et al. (2018b) for conditionally linear multiple-equation models.

Since, in many applications, the number of covariates in Eq. 1 is large, a dimensionality reduction strategy is needed. In this paper, we follow Papadogeorgou et al. (2021) and Billio et al. (2023, 2024) and consider a low-rank representation combined with a hierarchical prior distribution. The hierarchical prior allows for shrinking effects in the coefficient matrices $B_{\ell k}$, $k = 1, \dots, K$, and the low-rank representation induces further shrinking effects along different modes. We assume a PARAFAC representation and decompose the state-specific coefficient matrix as follows

$$B_{\ell k} = \sum_{d=1}^D B_{\ell,1,k}^{(d)} \circ B_{\ell,2,k}^{(d)} \quad (3)$$

where \circ is the Hadamard product and $B_{\ell,j,k}^{(d)}$ for $j = 1, 2$ are multiplicative factors (Papadogeorgou et al. (2021)). To simplify exposition, we assume there are two modes, $m = 2$, and the number of elements in each mode is p_1 and p_2 , respectively.

The hierarchical prior distribution includes three stages. At the first stage a matrix-variate normal distribution (Gupta and Nagar, 1999) is assumed

$$B_{\ell,m,k}^{(d)} \sim \mathcal{MN}_{p_1,p_2} \left(G_{\ell,m,k}^{(d)}, \tau_{\ell,k} \sigma_{\ell,m,k}^2 \zeta_{\ell,k}^{(d)} I_{p_1}, I_{p_2} \right) \quad (4)$$

where $\mathcal{MN}_{p_1,p_2}(M, U, V)$ denotes the matrix-variate normal distribution with $p_1 \times p_2$ mean matrix M , $p_1 \times p_1$ row covariance matrix U and $p_2 \times p_2$ column covariance matrix V . The location matrix $G_{\ell,m,k}^{(d)}$ is parametrized as follows:

$$G_{\ell,m,k}^{(d)} = \begin{cases} \gamma_{\ell,1}^{(d)} \otimes \boldsymbol{\nu}_{p_2}, & \text{if } m = 1 \\ \boldsymbol{\nu}_{p_1} \otimes \gamma_{\ell,2}^{(d)}, & \text{if } m = 2 \end{cases}$$

where \otimes denotes the *outer product*, $\boldsymbol{\nu}_n = (1, \dots, 1)'$ is the n -dimensional unit vector, $\gamma_{\ell,1}^{(d)}$ and $\gamma_{\ell,2}^{(d)}$ are the PARAFAC marginals in the conditional mean of the factors $B_{\ell,k}^{(d)}$, such that

$$\mathbb{E} \left(B_{\ell,k}^{(d)} \mid \gamma_{\ell,1,k}^{(d)}, \gamma_{\ell,2,k}^{(d)} \right) = \sum_{d=1}^D \mathbb{E} \left(B_{\ell,1,k}^{(d)} \right) \circ \mathbb{E} \left(B_{\ell,2,k}^{(d)} \right) = \sum_{d=1}^D \left(G_{\ell,1,k}^{(d)} \circ G_{\ell,2,k}^{(d)} \right) = \sum_{d=1}^D \gamma_{\ell,1,k}^{(d)} \otimes \gamma_{\ell,2,k}^{(d)}$$

In the second stage, we assume that the marginals from the PARAFAC decomposition follow a multivariate normal distribution

$$\boldsymbol{\gamma}_{\ell,m,k}^{(d)} \sim \mathcal{N}_{p_{m,k}}(\mathbf{0}, \tau_{\ell,k} \zeta_{\ell,k}^{(d)} W_{\ell,m,k}^{(d)}) \quad (5)$$

and assume the distributions are centred around the null vector and have random scales to allow for shrinkage at different levels. At the third stage, we borrow from Guhaniyogi et al. (2017) and specify the priors for the scales to induce shrinkage across components and rows

$$\tau_{\ell,k} \sim \mathcal{G}a(a_\tau, b_\tau) \quad (6)$$

$$\sigma_{\ell,m,k}^2 \sim \mathcal{G}a(a_\sigma, b_\sigma) \quad (7)$$

$$w_{\ell,m,j_m,k}^{(d)} \sim \text{Exp}((\lambda_{\ell,m,k}^{(d)})^2/2) \quad (8)$$

$$\lambda_{\ell,m,k}^{(d)} \sim \mathcal{G}a(a_\lambda, b_\lambda) \quad (9)$$

$$(\zeta_{\ell,k}^{(1)}, \dots, \zeta_{\ell,k}^{(D)}) \sim \text{Dir}(\alpha/D, \dots, \alpha/D) \quad (10)$$

where $\tau_{\ell,k}$ is a global scale which contributes to the variances of both $\boldsymbol{\gamma}_{\ell,m,k}^{(d)}$ and $B_{\ell,m,k}^{(d)}$. The matrices $W_{\ell,m,k}^{(d)} = \text{diag}(w_{\ell,1,1,k}^{(d)}, \dots, w_{\ell,m,j_m,k}^{(d)}, \dots, w_{\ell,M,p_M,k}^{(d)})$ are the row-specific parameters that shrink the individual elements of the marginals and, together with the prior on $\lambda_{\ell,m,k}^{(d)}$, lead to an adaptive LASSO-type penalty on $\boldsymbol{\gamma}_{\ell,m,k}^{(d)}$ (Armagan et al., 2013). The parameter $\zeta_{\ell,k}^{(d)}$ is component-specific and allows a subset of the D components to contribute substantially to the PARAFAC approximation while leaving the values of other components close to zero.

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 ℓ th coefficient tensor \mathbf{B}_ℓ to elicit the default choice of hyperparameters. In particular, the variance of the entry of coefficient tensor \mathbf{B}_ℓ for the soft PARAFAC can be written as a function of the hyperparameters:

$$\mathbb{V}(B_{\ell,ij}) = \frac{a_\tau(a_\tau + 1)}{b_\tau^2} C \left(\frac{a_\sigma}{b_\sigma} + \frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} \right)^2 \quad (11)$$

where $C = \frac{\alpha+1}{\alpha+1}$. Moreover, the variance of the coefficient entries for the hard PARAFAC is:

$$\mathbb{V}^{\text{hard}}(B_{\ell,ij}) = \frac{a_\tau(a_\tau + 1)}{b_\tau^2} C \left(\frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} \right)^2$$

We define the *relative additional variance* introduced by the softening of the PARAFAC as:

$$AV = \frac{\mathbb{V}(B_{\ell,ij}) - \mathbb{V}^{\text{hard}}(B_{\ell,ij})}{\mathbb{V}(B_{\ell,ij})}$$

which can be used to elicit the choice of the hyper-parameters.

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

$$\frac{a_\sigma}{b_\sigma} = \frac{b_\tau}{a_\tau} \sqrt{\frac{a_\tau V^*}{(a_\tau + 1)C}} \left(1 - \sqrt{1 - AV^*} \right) \quad (12)$$

Equation (11) and Proposition 1 are used in the simulations and empirical applications to help choose hyperparameters. In particular, we impose restrictions on the induced prior variance such that $\mathbb{V}(B_{\ell,ij}) = 1$ and $AV = 10\%$. Moreover, we set $\alpha = 1, a_\tau = 3, a_\sigma = 0.5, a_\lambda = 3, b_\lambda = \sqrt[2M]{a_\lambda}$ following Papadogeorgou et al. (2021) and we compute the values of b_τ and b_σ from Equations (11) and (12) 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 demonstrate through simulation studies (Appendix C in the Supplement) that our proposed MCMC procedure is robust to the selection of D , where true tensor coefficients with low to full ranks are used for data generation in different simulation settings. The inference performance are similar when $D \in \{3, 5, 7\}$. Moreover, additional simulations are carried out as a robustness check in which the true coefficients are contaminated with white noise such that the ranks are considered full for all different coefficients, and the MCMC procedure can recover the patterns of the true coefficients reasonably well for all values of $D \in \{3, 5, 7\}$ (Figure C.5 in Appendix C).

Alternatively, one can resort to model selection methods to choose the best value for D , since the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) can be computed for models with different values of D .

3 Posterior Approximation

In this section, we assume tensor-valued covariates and denote with $B_{\ell,m,\tilde{j}_m,k}^{(d)}$ the j_m^{th} slice of tensor $B_{\ell,m,k}^{(d)}$ along mode m , where $\tilde{j}_m = (:, :, \dots, j_m, \dots, :)$ and with $B_{\ell,m,\tilde{j}_m,k}^{(d)}$ the $p_1 \times \dots \times p_{m-1} \times p_{m+1} \times \dots \times p_M$ tensor with $m - 1$ modes. Let $\beta_{\ell,m,j_m,k}^{(d)} = \text{vec}(B_{\ell,m,\tilde{j}_m,k}^{(d)})$ be a $q_m \times 1$ vector, where $q_m = \prod_{l \neq m} p_l$. We further define the collections $\beta_k = (\beta_{1k}, \dots, \beta_{Nk})$ and $\gamma_k = (\gamma_{1k}, \dots, \gamma_{Nk})$, with $\beta_{\ell,k} = (\beta_{\ell,1,k}^{(1)}, \dots, \beta_{\ell,m,j_m,k}^{(d)}, \dots, \beta_{\ell,M,p_M,k}^{(D)})'$ and $\gamma_{\ell,k} = (\gamma_{\ell,1,k}^{(1)}, \dots, \gamma_{\ell,m,j_m,k}^{(d)}, \dots, \gamma_{\ell,M,p_M,k}^{(D)})'$, for $l = 1, \dots, N, k = 1, \dots, K, m = 1, \dots, M, d = 1, \dots, D$ and $j_m = 1, \dots, p_m$.

We summarize our Bayesian model in the Directed Acyclic Graph (DAG) representation of Fig. 1 where $\mathbf{y}_t = (y_{1t}, \dots, y_{Nt})$ is the collection of response variables across equations, $\mathbf{p} = (p_{11}, \dots, p_{1k}, \dots, p_{k1}, \dots, p_{kk})$ is the collection of transition probabilities, and $\beta = (\beta_1, \dots, \beta_K)$ and $\gamma = (\gamma_1, \dots, \gamma_K)$ denotes the collections across equations and states of the regression coefficients and PARFAC factors, respectively.

In the same diagram $\zeta = (\zeta_1, \dots, \zeta_K)$, $\tau = (\tau_1, \dots, \tau_K)'$, $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$ and $\lambda = (\lambda_1, \dots, \lambda_K)$ denote the collections of the hyper-parameters at the second and third stage of the hierarchical prior where $\zeta_k = (\zeta_{1,k}, \dots, \zeta_{N,k})$, $\zeta_{\ell,k} = (\zeta_{\ell,k}^{(1)}, \dots, \zeta_{\ell,k}^{(d)}, \dots, \zeta_{\ell,k}^{(D)})'$, $\lambda_k = (\lambda_{1k}, \dots, \lambda_{Nk})$, $\lambda_{\ell,k} = (\lambda_{\ell,1,k}^{(1)}, \dots, \lambda_{\ell,m,k}^{(d)}, \dots, \lambda_{\ell,M,k}^{(D)})'$ and $\mathbf{w}_k = (\mathbf{w}_{1k}, \dots, \mathbf{w}_{Nk})$, $\mathbf{w}_{\ell,k} = (w_{\ell,1,k}^{(1)}, \dots, w_{\ell,m,j_m,k}^{(d)}, \dots, w_{\ell,M,p_M,k}^{(D)})'$.

3.1 Sampling method

The joint posterior distribution is not tractable, so we develop a MCMC algorithm to sample from it. Specifically, we use a Gibbs sampling procedure which combines two sampling strategies: i) back-fitting sampling (Hastie and Tibshirani, 2000) for the coefficients and ii) forward filtering and backward sampling for the latent states (Frühwirth-Schnatter, 2006). To cope with the computational cost of the Monte Carlo approximation, we implement a version of random scan Gibbs Łatuszyński et al. (2013).

Let $\theta = (\theta_1, \dots, \theta_K)$ be the collection of the state-specific parameters $\theta_k = (\beta_k, \gamma_k, \zeta_k, \tau_k, \lambda_k, \mathbf{w}_k)$ and denote with $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_T)$, $\mathbf{X} = (X_1, \dots, X_T)$ and $\mathbf{s} = (s_1, \dots, s_T)$ the collection of response variables, covariates and state variables, respectively.

Since the joint posterior $p(\theta|\mathbf{y}, \mathbf{X})$ is not tractable, we follow a data augmentation strategy and introduce the joint posterior $p(\theta, \mathbf{s}|\mathbf{y}, \mathbf{X})$. 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 Papadogeorgou et al. (2021). First, we include the global shrinkage

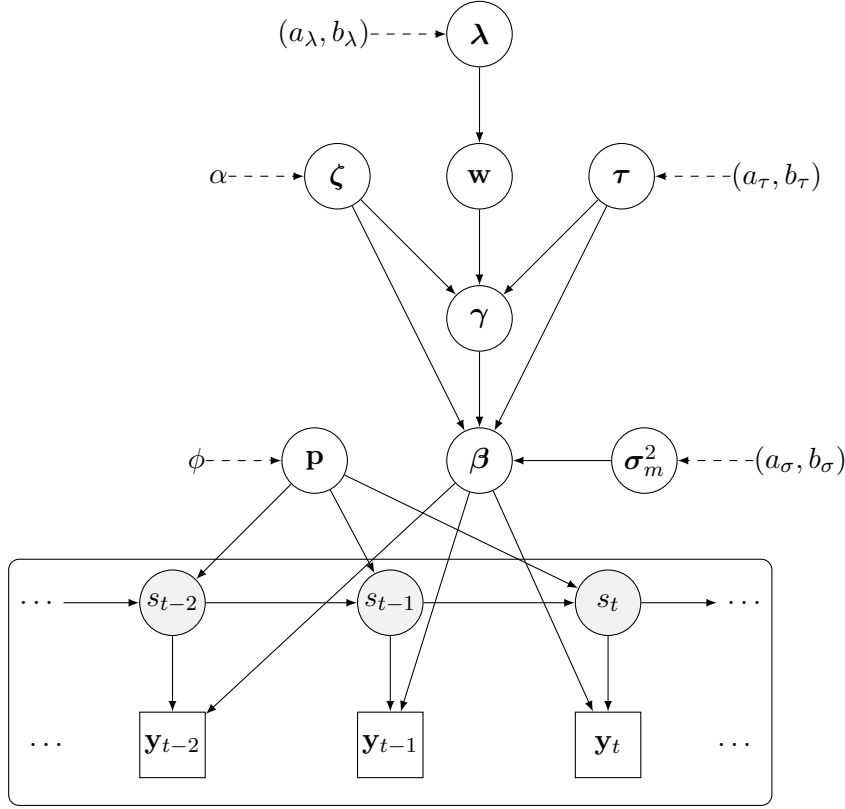


Figure 1: DAG of the Bayesian Markov-switching Matrix Regression model. It exhibits the hierarchical structure of the observations \mathbf{y}_t (boxes), the latent state variables s_t (grey circles), the parameters $\beta_{\ell,m,j_m,k}^{(d)}$, $\sigma_{\ell,k}^2$, the hyper-parameters of the first stage $\gamma_{\ell,m,j_m,k}^{(d)}$, the second stage $\tau_{\ell,k}$, $\zeta_{\ell,m,k}^{(d)}$ and $w_{\ell,m,j_m,k}^{(d)}$ and the third stage $\lambda_{\ell,m,k}^{(d)}$ (white circles). The directed arrows show the conditional independence structure of the model.

parameter τ not only in the prior for γ but also in the prior for β . Second, we integrate out γ from the full conditional of β to allow β to depend directly on the observed data. The resulting collapsed Gibbs sampler allows us to achieve exact sampling for β and γ and to improve the sampler efficiency (Robert et al., 2007). Third, we apply random scan Gibbs to increase the efficiency of the sampler (Łatuszyński et al., 2013).

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. Its implementation relies on the following equivalent representation of the regression model.

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

$$y_{\ell t} = \sum_{k=1}^K \left(\beta_{\ell,m,j_m,k}^{(d)\top} \Psi_{\ell,j_m,m,t,k}^{(d)} + R_{\ell,j_m,m,t,k}^{(d)} + R_{\ell,t,k}^{(d)} \right) \mathbb{I}(s_t = k) + \sum_{k=1}^K \sigma_{\ell,k}^2 \varepsilon_{\ell,t} \mathbb{I}(s_t = k)$$

$\ell = 1, \dots, N$, where the residual terms $R_{\ell,dt,k}$ and $R_{\ell,j_mdt,k}$ and the auxiliary covariate vector Ψ_{ℓ,j_mdt} are:

$$\begin{aligned} R_{\ell,t}^{(d)}(s_t) &= \sum_{d' \neq d} \langle B_{\ell,1}(s_t)^{(d')} \circ \dots \circ B_{\ell,M}(s_t), X_t \rangle \\ R_{\ell,m,j_m,t}^{(d)}(s_t) &= \langle (B_{\ell,1}^{(d)}(s_t) \circ \dots \circ B_{\ell,M}^{(d)}(s_t))_{-j_m}, (X_t)_{-j_m} \rangle \\ \Psi_{\ell,m,j_m,t}^{(d)}(s_t) &= \text{vec}((B_{\ell,1}^{(d)}(s_t) \circ \dots \circ B_{\ell,m-1}^{(d)}(s_t) \circ B_{\ell,m+1}^{(d)}(s_t) \circ \dots \circ B_{\ell,M}^{(d)}(s_t) \circ X_t)_{\bar{j}_m}) \end{aligned}$$

where $(A)_{-j_m}$ indicates the tensor after removing the j_m -th slice along mode m and $\tilde{j}_m = (:, \dots, :, j_m, :, \dots, :)$ is the collection of indexes along $m - 1$ modes while keeping fix the index j_m of the mode m .

At every iteration of the sampling algorithm, we randomly select the factor index d from the proposal distribution $g(d)$ with support $\{1, 2, \dots, D\}$ and the model index m from the distribution $h(m)$ with support $\{1, 2, \dots, M\}$. For $k = 1, \dots, K$ and $\ell = 1, \dots, N$, all the elements of $B_{\ell,k}$ and the PARAFAC margins $\gamma_{\ell,m,j_m,k}^{(d)}$ are sampled from their full conditional distributions:

1. Draw $\beta_{\ell,m,j_m,k}^{(d)}$ from $f(\beta_{\ell,m,j_m,k}^{(d)} | \mathbf{y}, B, \beta_{\ell,m,-j_m,k}^{(d)}, \zeta_{\ell,k}^{(d)}, \tau_{\ell,k}, \lambda_{\ell,m,k}^{(d)}, \mathbf{w}_{\ell,m,j_m,k}^{(d)})$ for $d \in \{1, \dots, D\}$ which is a multivariate normal distribution, where the d and m have been randomly selected, i.e. $d \sim g(d)$ and $m \sim h(m)$.
2. Draw $\gamma_{\ell,m,j_m,k}^{(d)}$ from $f(\gamma_{\ell,m,j_m,k}^{(d)} | \beta, \zeta_{\ell,k}^{(d)}, \tau_{\ell,k}, \lambda_{\ell,m,k}^{(d)}, \mathbf{w}_{\ell,m,j_m,k}^{(d)})$, which is a univariate normal distribution.

The remaining parameters and hyper-parameters are sampled via the following Gibbs updates:

3. Draw $\zeta_{\ell,k}$ from the Generalized Inverse Gaussian distribution $f(\zeta_{\ell,k} | \beta_k, \gamma_{\ell,k}, \lambda_{\ell,m,k}^{(d)}, \mathbf{w}_{\ell,m,j_m,k}^{(d)})$.
4. Draw $\tau_{\ell,k}$ from the Generalized Inverse Gaussian distribution $f(\tau_{\ell,k} | \beta_{\ell,k}, \gamma_{\ell,k}, \zeta_{\ell,k}^{(d)}, \mathbf{w}_{\ell,k}^{(d)})$.
5. Draw $\lambda_{\ell,k}$ from $f(\lambda_{\ell,k} | \gamma_{\ell,k}, \tau_{\ell,k}, \zeta_{\ell,k})$ which is a Gamma distribution.
6. Draw $\mathbf{w}_{\ell,k}$ from the Generalized Inverse Gaussian distribution $f(\mathbf{w}_{\ell,k} | \gamma_{\ell,k}, \lambda_{\ell,k}, \tau_{\ell,k}, \zeta_{\ell,k})$.
7. Draw transitional probabilities (p_{k1}, \dots, p_{kK}) from Dirichlet distribution $f(p_{k1}, \dots, p_{kK} | \mathbf{s})$.

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

8. Compute iteratively the vector of smoothed probabilities $\xi_{t|T} = p(s_t | \boldsymbol{\theta}, \mathbf{y})$ 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 $\boldsymbol{\theta} \in R^d$, but at each iteration, only a random subset of fixed size, say $k \leq d$, of the parameter vector, is updated. Moreover, every set of indices of size k 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 (PRSG).

Algorithm 1 The steps in a Random-Partial-Scan Gibbs

- S1: Draw uniformly $I \subset \{1, \dots, d\}$ a random set of indices of size $k \leq d$ so that each subset has an equal chance of being selected.
 - S2: If $I = (i_1, \dots, i_k)$, update $\theta_I = (\theta_{i_1}, \theta_{i_2}, \dots, \theta_{i_k})$ using a random scan and leave the other components of θ unchanged.
-

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

Proposition 3. *The chain generated by the PRSG 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. 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 estimates of the coefficients, and the regimes are successfully identified in the different experimental settings (see Appendix C in the Supplementary for further details).

4 Empirical Application

We test the validity of our tensor regression model using two real-world applications. In particular, we would like to show through the applications that our tensor regression model with Markov Switching: i) outperforms the competing models (ordinary least square 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 Markov Switching Model.

4.1 Volatility Index of 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 oil volatility index (OVX) with a number of 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, Fernandes et al. (2014) studied the long-range dependence in the VIX data by including a vector of the average of the logarithm of VIX for the last $k \in \{1, 5, 10, 22, 66\}$ days (to mirror daily, weekly, bi-weekly, monthly and quarterly component) in a family of heterogeneous autoregressive (HAR) processes.

We follow similar strategy as Fernandes et al. (2014) in forecasting VIX, but adapt it into a multiple-equation tensor regression framework, where we regress VIX on OVX (Eq. (13)) and vice versa (Eq. (14)) together with other covariates: the k day log return for S&P 500, exchange rate (proxy by US dollar index), spot price of WTI crude oil for $k \in \{1, \dots, 44\}$. 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 is shown in Equation (13) and Equation (14).

$$\text{VIX}_t = \mu_1(s_t) + \left\langle B_1(s_t), \begin{pmatrix} \text{SP}_{t-1} & \dots & \text{SP}_{t-h} & \dots & \text{SP}_{t-44} \\ \text{ER}_{t-1} & \dots & \text{ER}_{t-h} & \dots & \text{ER}_{t-44} \\ \text{Oil}_{t-1} & \dots & \text{Oil}_{t-h} & \dots & \text{Oil}_{t-44} \\ \text{OVX}_{t-1} & \dots & \text{OVX}_{t-h} & \dots & \text{OVX}_{t-44} \end{pmatrix} \right\rangle + \sigma_1(s_t)\epsilon_{1t} \quad (13)$$

$$\text{OVX}_t = \mu_2(s_t) + \left\langle B_2(s_t), \begin{pmatrix} \text{SP}_{t-1} & \dots & \text{SP}_{t-h} & \dots & \text{SP}_{t-44} \\ \text{ER}_{t-1} & \dots & \text{ER}_{t-h} & \dots & \text{ER}_{t-44} \\ \text{Oil}_{t-1} & \dots & \text{Oil}_{t-h} & \dots & \text{Oil}_{t-44} \\ \text{VIX}_{t-1} & \dots & \text{VIX}_{t-h} & \dots & \text{VIX}_{t-44} \end{pmatrix} \right\rangle + \sigma_2(s_t)\epsilon_{2t} \quad (14)$$

We show the in-sample fitting results of the MSMETR against the competing models in Figure 2. The in-sample fitting of the OLS and Linear LASSO regression fails to capture the structural changes in the series of VIX and OVX. However, these structural changes are successfully captured by an MSMETR, for which we assumed there are two possible regimes representing high and low levels of volatility. Indeed, the MSMETR identified two different regimes with distinctive regime-specific intercepts. Regime 2, representing a high level of volatility, has a higher intercept value than Regime 1, which represents a low level of volatility.

We show the effects of k -day log return of oil prices and S&P 500 on VIX (blue dots) and OVX (red dots), respectively, in Figure 3. 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 k -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 larger effect on OVX than on VIX in the high-volatility regime. Oil returns have a small impact on VIX and a more pronounced impact on OVX.

There is also strong evidence of non-negligible long-term effects of oil prices (dark red points) on oil volatility, consistently with previous findings (Bandi and Perron, 2006; Corsi, 2009). The

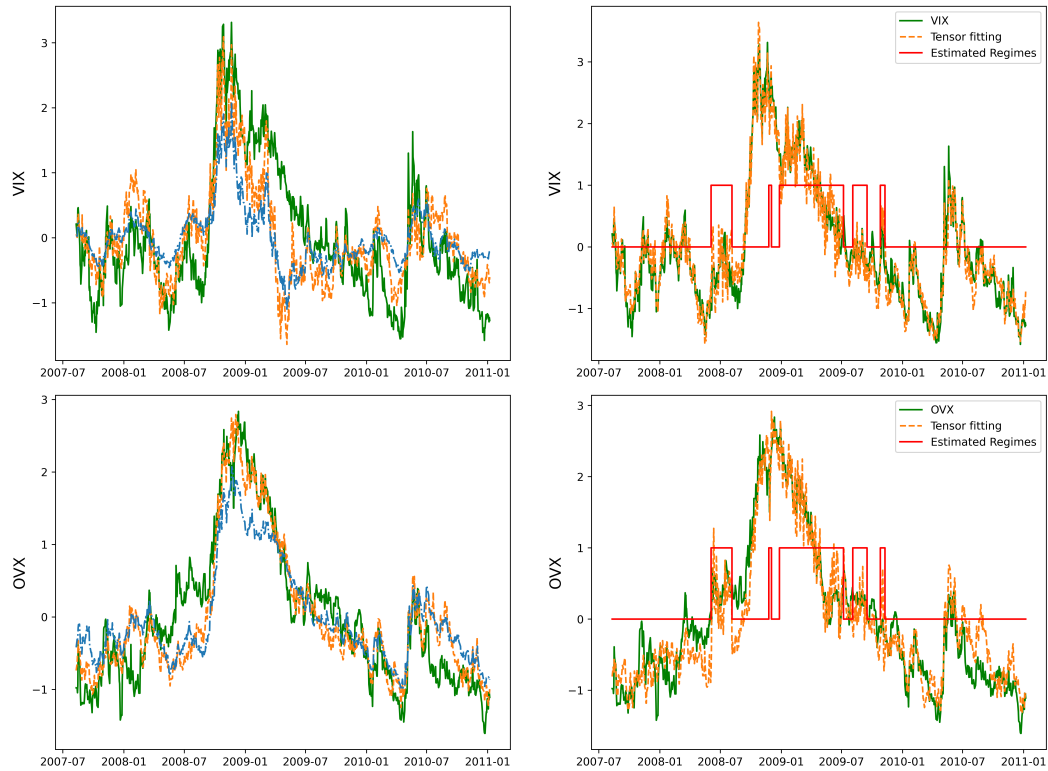


Figure 2: Left: In-sample fitting for Least Square (orange dashed) and LASSO (blue dashed). Right: In-sample fitting of Tensor Regression with Markov Switching (orange dashed) and estimated hidden states (red solid). The green solid line represents the VIX and VOX indexes (top and bottom).

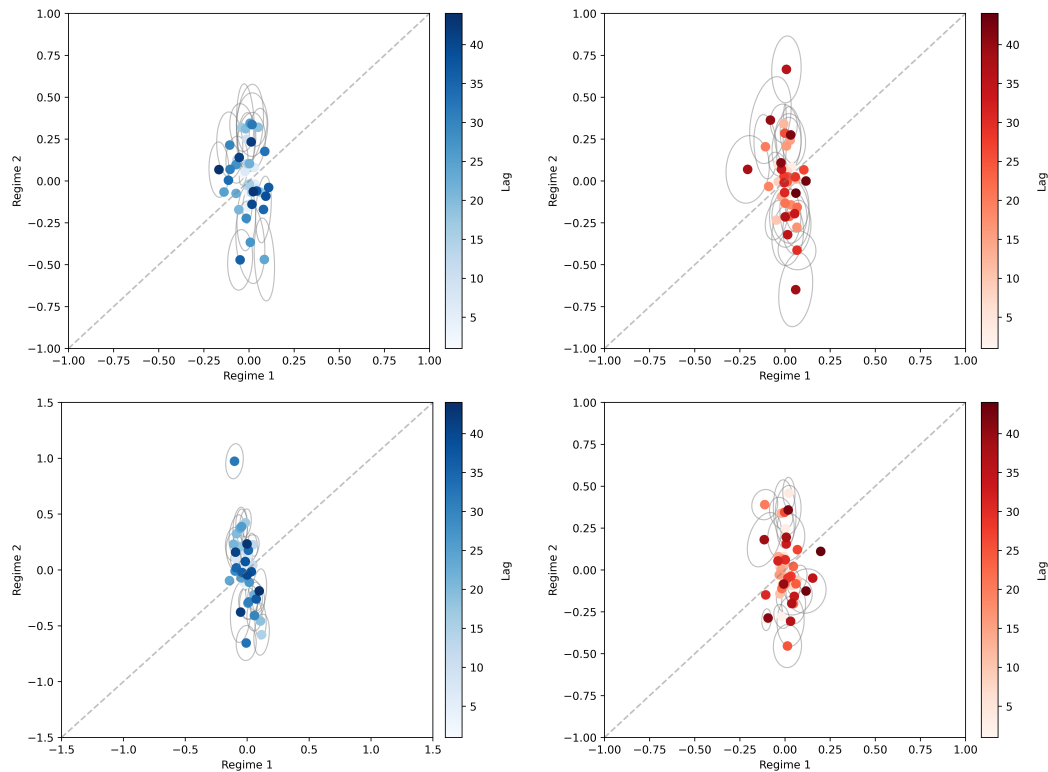


Figure 3: Effects of k -day Oil (top) and S&P 500 (bottom) log returns on VIX (left) and OVX (right) for $k \in \{1, \dots, 44\}$. Lighter and darker colors represent smaller and larger k , respectively. 90% Highest Posterior Density regions (gray ellipses) are plotted for the coefficients which exhibit asymmetric effects across regimes (HPD ellipse does not intersect the 45° line).

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 effects of S&P 500 on OVX follow 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.

In Table D.1 of the Supplement, we report the mean square error (MSE) and mean absolute error (MAE) for the in-sample fitting and out-of-sample forecasting with prediction horizons of 1-day and 5-day. Our proposed MSMETR consistently outperforms the Least Squares and linear LASSO across almost all the measures.

4.2 Oil Prices on Stock Return

For the second application, we extend our matrix-variate tensor regression model to a 3-mode tensor regression model by constructing the covariates \mathcal{X}_t as a three-dimensional array for each observation. Therefore, the coefficients \mathcal{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 (e.g., see Xiao and Wang, 2022; Xiao et al., 2023) and examine the impact of oil price volatility on the stock market returns (S&P 500) at an aggregate level and on 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 Xiao and Wang (2022) in that we consider a Mixed Data Sampling (MIDAS) (Ghysels et al., 2004) 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 $R_{\ell,t}$ is the 4-week log-return of market ℓ at time t , where $\ell = \{\text{S\&P 500, financial sector, energy sector, other sectors in S\&P 500}\}$. The covariates are sampled weekly at the 1st week, 2nd week, 3rd week, 4th week before time t , indexed by $t - \frac{1}{4}, t - \frac{2}{4}, t - \frac{3}{4}, t - \frac{4}{4}$. 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 Xiao and Wang (2022).

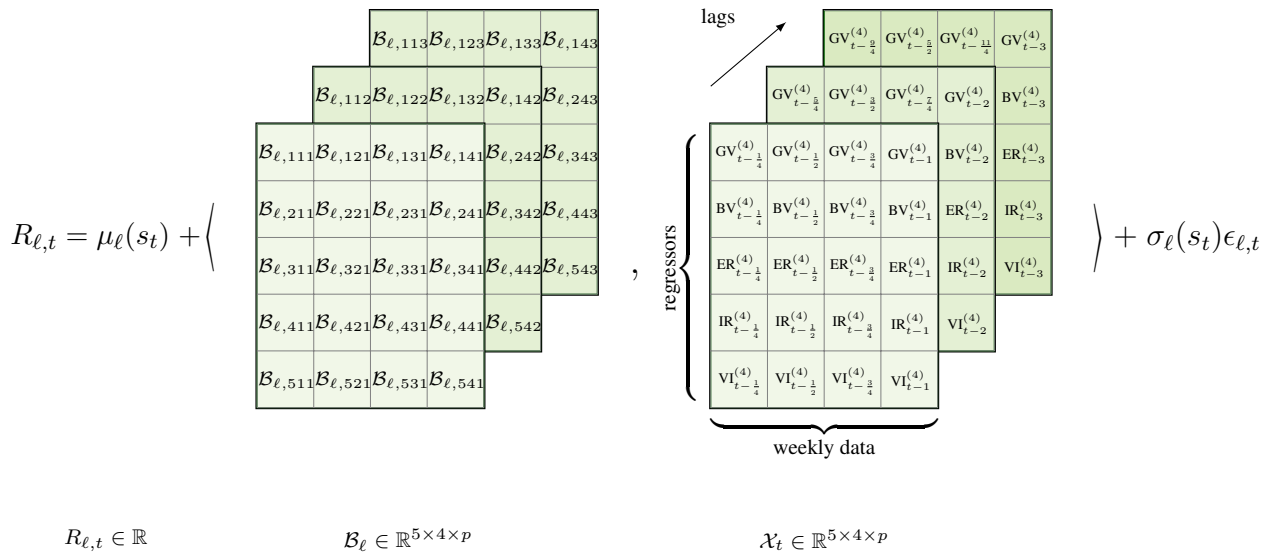


Figure 4: Graphic Representation of Tensor Regression for Macro Application

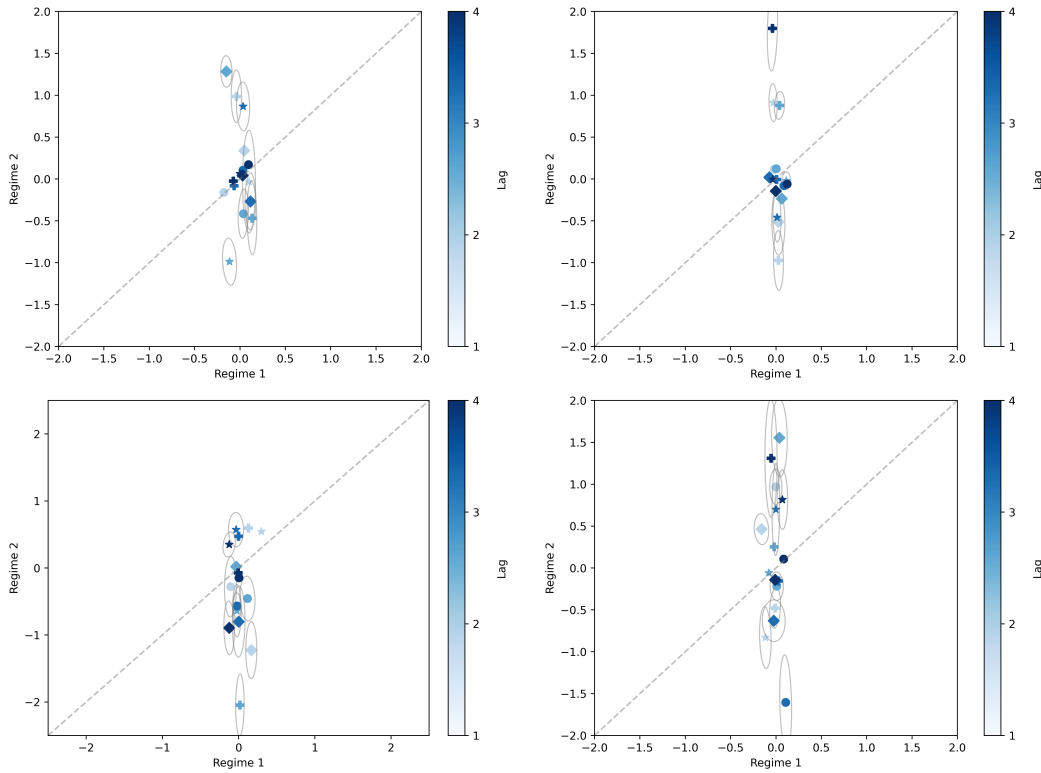


Figure 5: The scatter plot shows the effects of Good Oil Volatility (left) and Bad Oil Volatility (right) on the financial sector (top) and energy sector (bottom). Different symbols represent different weeks, with \bullet : $t - (1 + 4(p - 1))/4$, \blacksquare : $t - 2(1 + 4(p - 1))/4$, \blacklozenge : $t - 3(1 + 4(p - 1))/4$ and \blackstar : $t - 4(1 + 4(p - 1))/4$ for $p = \{1, 2, 3, 4\}$. Different blue shades represent lags, from order 1 (lighter) to order 4 (darker).

We arrange the different regressors along the rows (first mode) of the tensor covariates, data points sampled weekly for four weeks along the columns (second mode), and the weekly data points of the previous four weeks are stacked along the third axis (third mode) for p lags. A graphical representation of the specification of the tensor regression model is shown in Fig. 4, where $\mu_\ell(s_t)$ and $\sigma_\ell(s_t)$ are regime-specific intercepts and standard deviations, \mathcal{B}_ℓ and \mathcal{X}_t are, respectively, tensor coefficients and covariates of sizes $\mathbb{R}^{5 \times 4 \times p}$, and p represents the number of lags we wish to consider for the covariates.

Figure D.1 of the Supplement shows the in-sample fitting of Least Squares and Linear LASSO (left column) and the in-sample fitting of MSMETR (right column). Notably, Least Squares and linear LASSO fail to capture the volatility changes in the market return. In contrast, the MSMETR 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, MSMETR 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 Figure D.1 of the Supplement), when sector indices are used as dependent variables, MSMETR 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 MSMETR can capture more structural changes at a disaggregated level can be largely attributed to the heterogeneity between different sectors. Thus, MSMETR with multiple equations can also be an effective data integration tool.

Figure 5 shows the effects of GV and BV on financial and energy sector log-returns (see also Figure D.2 of the Supplement). We use different symbols to represent the weekly data sampled at different weeks for different lags $p = \{1, 2, 3, 4\}$, with \bullet : $t - (1 + 4(p - 1))/4$, \blacksquare : $t - 2(1 + 4(p - 1))/4$, \blacklozenge : $t - 3(1 + 4(p - 1))/4$ and \blackstar : $t - 4(1 + 4(p - 1))/4$. And we use lighter (darker) blue to represent lower (higher) lag p . 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 (Xiao and Wang, 2022; Xiao et al., 2023). 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 larger 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 Xiao et al. (2023).

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 MSMETR 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 to work with high dimensional data where a common hidden Markov chain process introduces dependencies 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 MSMETR with two real-world applications, where MSMETR outperforms the competing algorithms in both in-sample fitting and out-of-sample forecasting. Moreover, MSMETR 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 MSMETR can also capture the heterogeneity in the data at aggregate and disaggregate levels to exploit more information in the data.

The model we proposed is ready to be used for tensor regression with order 2 or 3. It can be applied to many other applications where regression on high-dimensional data is needed, and over-parametrization or overfitting issues must be considered.

Funding

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 of the Results

A.1 Proof of Proposition 1

In the following, we drop the state subscript index k and equation subscript index ℓ for simplicity. The variance of the coefficient entries of soft PARAFAC can be written as:

$$\mathbb{V}(B_{ij}) = \mathbb{E} \{ \mathbb{V}(B_{ij} | S, Z, W, \tau) \} = \mathbb{E} \left\{ \text{var} \left(\sum_{d=1}^D \prod_{m=1}^M \beta_{m,(ij)}^{(d)} | S, Z, W, \tau \right) \right\}$$

$$\begin{aligned}
&= \mathbb{E} \left\{ \sum_{d=1}^D \prod_{m=1}^M \text{Var} \left(\beta_{m,(ij)}^{(d)} | S, Z, W, \tau \right) \right\} = \mathbb{E} \left\{ \sum_{d=1}^D \prod_{m=1}^M \tau \zeta^{(d)} \left(\sigma_m^2 + w_{m,(ij)}^{(d)} \right) \right\} \\
&= \mathbb{E} \{ \tau^M \} \mathbb{E} \left\{ \sum_{d=1}^D \left(\zeta^{(d)} \right)^M \right\} \mathbb{E} \left\{ \prod_{m=1}^M \left(\sigma_m^2 + w_{m,(ij)}^{(d)} \right) \right\} \\
&= \frac{\Gamma(a_\tau + M)}{\Gamma(a_\tau) b_\tau^M} D \prod_{r=0}^{M-1} \frac{\alpha/D + r}{\alpha + r} \left(\frac{a_\sigma}{b_\sigma} + \frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} \right)^M \tag{A.1}
\end{aligned}$$

For the variance of the coefficient entries of hard PARAFAC, $\sigma_m^2 = 0$, thus

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

It's not hard to notice that $\frac{a_\sigma}{b_\sigma}$ is the quantity that drives the additional variability of the soft PARAFAC by comparing Equations (A.1) and (A.2). The goal is to set $\mathbb{V}(B_{ij}) = V^*$ and $AV = AV^*$. By exploiting $\mathbb{V}(B_{ij})/\mathbb{V}^{\text{hard}}(B_{ij}) = (1 - AV^*)^{-1}$ and setting $M = 2$, we have

$$\frac{\mathbb{V}(B_{ij})}{\mathbb{V}^{\text{hard}}(B_{ij})} = \frac{\left(\frac{a_\sigma}{b_\sigma} + \frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} \right)^2}{\left(\frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} \right)^2} = \left(\frac{a_\sigma}{b_\sigma} \frac{(a_\lambda - 1)(a_\lambda - 2)}{2b_\lambda^2} + 1 \right)^2 = (1 - AV^*)^{-1}$$

Solving the above equation for $\frac{a_\sigma}{b_\sigma}$ and given $\frac{a_\sigma}{b_\sigma}$ is positive,

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

By setting $\mathbb{V}(B_{ij}) = V^*$ we have,

$$\frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} = \frac{b_\tau}{a_\tau} \sqrt{\frac{a_\tau V^*}{(a_\tau + 1)C}} - \frac{a_\sigma}{b_\sigma} \tag{A.4}$$

Combing Equations (A.3) and (A.4),

$$\frac{a_\sigma}{b_\sigma} = \frac{b_\tau}{a_\tau} \sqrt{\frac{a_\tau V^*}{(a_\tau + 1)C}} \left(1 - \sqrt{1 - AV^*} \right)$$

A.2 Proof of Proposition 2

For the likelihood part, we drop the state subscript index k and equation subscript index ℓ for simplicity. We have

$$\begin{aligned}
\langle B, X_t \rangle &= \left\langle \sum_{d=1}^D B_1^{(d)} \circ \dots \circ B_M^{(d)}, X_t \right\rangle \\
&= \left\langle B_m^{(d)}, B_1^{(d)} \circ \dots \circ B_{m-1}^{(d)} \circ B_{m+1}^{(d)} \circ \dots \circ B_M^{(d)} \circ X_t \right\rangle + \sum_{d' \neq d} \left\langle B_1^{(d')} \circ \dots \circ B_M^{(d')}, X_t \right\rangle \\
&= \sum_{j_m=1}^{p_m} \left\langle B_{m,j_m}^{(d)}, (B_1^{(d)} \circ \dots \circ B_{m-1}^{(d)} \circ B_{m+1}^{(d)} \circ \dots \circ B_M^{(d)} \circ X_t)_{\tilde{j}_m} \right\rangle + R_t^{(d)} \\
&= \beta_{m,j_m}^{(d)'} \text{vec}(B_1^{(d)} \circ \dots \circ B_{m-1}^{(d)} \circ B_{m+1}^{(d)} \circ \dots \circ B_M^{(d)} \circ X_t)_{\tilde{j}_m} + R_{j_m,m,t}^{(d)} + R_t^{(d)} \\
&= \beta_{m,j_m}^{(d)'} \Psi_{j_m,m,t}^{(d)} + R_{j_m,m,t}^{(d)} + R_t^{(d)}
\end{aligned}$$

where

$$\begin{aligned}
R_t^{(d)} &= \sum_{d' \neq d} \langle B_1^{(d')} \circ \dots \circ B_M^{(d')}, X_t \rangle \\
R_{j_m, m, t}^{(d)} &= \sum_{j'_m \neq j_m} \langle B_{m, j'_m}^{(d)}, (B_1^{(d)} \circ \dots \circ B_{m-1}^{(d)} \circ B_{m+1}^{(d)} \circ \dots \circ B_M^{(d)} \circ X_t)_{j'_m} \rangle \\
&= \sum_{j'_m \neq j_m} \langle (B_1^{(d)} \circ \dots \circ B_M^{(d)})_{j'_m}, (X_t)_{j'_m} \rangle \\
&= \langle (B_1^{(d)} \circ \dots \circ B_M^{(d)})_{-j_m}, (X_t)_{-j_m} \rangle \\
\Psi_{j_m, m, t}^{(d)} &= \text{vec}(B_1^{(d)} \circ \dots \circ B_{m-1}^{(d)} \circ B_{m+1}^{(d)} \circ \dots \circ B_M^{(d)} \circ X_t)_{j_m}
\end{aligned}$$

A.3 Proof of Proposition 3

To simplify notation, we prove the result for $|I|=3$, but the result holds in general. Suppose 2 components with indexes set $\mathbf{d} = \{d_1, d_2, d_3\}$ and $|\mathbf{d}| = 3$ are randomly selected to be updated, and the order in which the $|\mathbf{d}|$ components are updated is also random.

Assume that the set of indexes is selected at random and the order in which the components are updated is also random. Given that the other components outside of the index set \mathbf{d} remain unchanged, we have the transition function

$$\begin{aligned}
K(\theta^{(i+1)}|\theta^{(i)}) &= \xi_1 \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i)}, \theta_{d_3}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i+1)}, \theta_{-\mathbf{d}}^{(i)}) + \\
&\quad \xi_2 \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i)}, \theta_{d_3}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i+1)}, \theta_{-\mathbf{d}}^{(i)}) + \\
&\quad \xi_3 \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_3}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i+1)}, \theta_{-\mathbf{d}}^{(i)}) + \\
&\quad \xi_4 \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_3}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_1}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i+1)}, \theta_{-\mathbf{d}}^{(i)}) + \\
&\quad \xi_5 \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_2}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_3}^{(i+1)}, \theta_{d_2}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i+1)}, \theta_{-\mathbf{d}}^{(i)}) + \\
&\quad \xi_6 \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_2}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_3}^{(i+1)}, \theta_{d_1}^{(i)}, \theta_{-\mathbf{d}}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i+1)}, \theta_{-\mathbf{d}}^{(i)})
\end{aligned}$$

where $\xi_1 \sim \xi_6$ 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 $\theta_{-\mathbf{d}}^{(i)}$ from the equations for the rest of the proof. From the detailed balance condition, we have $K(\theta^{(i+1)}|\theta^{(i)}) \pi(\theta^{(i)}) = K(\theta^{(i)}|\theta^{(i+1)}) \pi(\theta^{(i+1)})$ which implies:

$$\begin{aligned}
&\left\{ \xi_1 \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i)}, \theta_{d_3}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i+1)}) + \right. \\
&\quad \xi_2 \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i)}, \theta_{d_3}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i+1)}) + \\
&\quad \xi_3 \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_3}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i+1)}) + \\
&\quad \xi_4 \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_3}^{(i)}) \pi(\theta_{d_3}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_1}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i+1)}) + \\
&\quad \xi_5 \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_2}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_3}^{(i+1)}, \theta_{d_2}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i+1)}) + \\
&\quad \left. \xi_6 \pi(\theta_{d_3}^{(i+1)}|\theta_{d_1}^{(i)}, \theta_{d_2}^{(i)}) \pi(\theta_{d_2}^{(i+1)}|\theta_{d_3}^{(i+1)}, \theta_{d_1}^{(i)}) \pi(\theta_{d_1}^{(i+1)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i+1)}) \right\} \pi(\theta^{(i)}) \\
&= \left\{ \xi_1 \pi(\theta_{d_1}^{(i)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i+1)}) \pi(\theta_{d_2}^{(i)}|\theta_{d_1}^{(i)}, \theta_{d_3}^{(i+1)}) \pi(\theta_{d_3}^{(i)}|\theta_{d_1}^{(i)}, \theta_{d_2}^{(i)}) + \right. \\
&\quad \xi_2 \pi(\theta_{d_1}^{(i)}|\theta_{d_2}^{(i+1)}, \theta_{d_3}^{(i+1)}) \pi(\theta_{d_3}^{(i)}|\theta_{d_1}^{(i)}, \theta_{d_2}^{(i+1)}) \pi(\theta_{d_2}^{(i)}|\theta_{d_1}^{(i)}, \theta_{d_3}^{(i)}) + \\
&\quad \xi_3 \pi(\theta_{d_2}^{(i)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i+1)}) \pi(\theta_{d_1}^{(i)}|\theta_{d_2}^{(i)}, \theta_{d_3}^{(i+1)}) \pi(\theta_{d_3}^{(i)}|\theta_{d_1}^{(i)}, \theta_{d_2}^{(i)}) + \\
&\quad \left. \xi_4 \pi(\theta_{d_2}^{(i)}|\theta_{d_1}^{(i+1)}, \theta_{d_3}^{(i+1)}) \pi(\theta_{d_3}^{(i)}|\theta_{d_2}^{(i)}, \theta_{d_1}^{(i+1)}) \pi(\theta_{d_1}^{(i)}|\theta_{d_2}^{(i)}, \theta_{d_3}^{(i)}) + \right. \\
&\quad \left. \xi_5 \pi(\theta_{d_3}^{(i)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i+1)}) \pi(\theta_{d_1}^{(i)}|\theta_{d_3}^{(i)}, \theta_{d_2}^{(i+1)}) \pi(\theta_{d_2}^{(i)}|\theta_{d_1}^{(i)}, \theta_{d_3}^{(i)}) + \right. \\
&\quad \left. \xi_6 \pi(\theta_{d_3}^{(i)}|\theta_{d_1}^{(i+1)}, \theta_{d_2}^{(i+1)}) \pi(\theta_{d_2}^{(i)}|\theta_{d_3}^{(i)}, \theta_{d_1}^{(i+1)}) \pi(\theta_{d_1}^{(i)}|\theta_{d_2}^{(i)}, \theta_{d_3}^{(i)}) \right\} \pi(\theta^{(i+1)})
\end{aligned}$$

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 (Frühwirth-Schnatter, 2006). Let us introduce the set of allocation variables $\boldsymbol{\xi}_t = (\xi_{1t}, \dots, \xi_{Kt})$, with $\xi_{kt} = \mathbf{I}(s_t = k)$. By means of dynamic factorization, the full conditional distribution of the hidden state is

$$p(s_1, \dots, s_t | \mathbf{y}, B, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\zeta}, \boldsymbol{\tau}, \mathbf{p}) \propto \prod_{t=1}^T \left(\prod_{\ell=1}^N p(y_{\ell,t} | B_{\ell}(s_t), \sigma_{\ell}^2(s_t)) \right) \prod_{k=1}^K \prod_{l=1}^K p_{lk}^{\xi_{kt}\xi_{l,t-1}} \quad (\text{B.1})$$

where $B = (B_1, \dots, B_K)$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)$, $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_K)$, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)'$.

B.2 Full conditional distribution of the transition probability

$$p((p_{1i}, \dots, p_{Ki}) | S, \mathbf{y}) \propto \prod_{t=1}^T \prod_{l=1}^K p_{li}^{\xi_{lt}\xi_{i,t-1}} \prod_{l=1}^K p_{li}^{\nu_l-1} \propto \prod_{l=1}^K p_{il}^{\sum_{t=1}^T \xi_{lt}\xi_{i,t-1} + \nu_l-1} \quad (\text{B.2})$$

which is proportional to Dirichlet distribution $\text{Dir}(\bar{\nu}_1, \dots, \bar{\nu}_K)$, where $\bar{\nu}_l = \nu_l + \sum_{t=1}^T \xi_{lt}\xi_{i,t-1}$.

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 $\boldsymbol{\beta}_{m,j_m}^{(d)} \sim \mathcal{N}_{q_m}(\boldsymbol{\gamma}_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m}, \tau \sigma_m^2 \zeta^{(d)} I_{q_m})$.

The posterior of the unknowns of the model is given by

$$p(\boldsymbol{\beta}_{m,j_m}^{(d)}, \boldsymbol{\gamma}_{m,j_m}^{(d)}, \sigma_m^2, \mathbf{w}_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)} | \mathbf{y}, B_1, \dots, B_T) \quad (\text{B.3})$$

We adopt the MCMC procedure based on the Gibbs sampling algorithm to sample the unknowns from 3 blocks.

Block 1: Sampling $\zeta^{(d)}$ and τ from $p(\zeta^{(d)}, \tau | B, \boldsymbol{\gamma}, \mathbf{w})$

We first sample ζ from the joint posterior by integrating out τ

$$\begin{aligned} p(\zeta | B, \boldsymbol{\gamma}, \mathbf{w}) &\propto \pi(\zeta) p(B, \boldsymbol{\gamma} | \mathbf{w}, \zeta^{(d)}) = \pi(\zeta) \int_{\tau} p(B | \boldsymbol{\gamma}, \mathbf{w}, \tau) p(\boldsymbol{\gamma} | \mathbf{w}, \tau) \pi(\tau) d\tau \\ &= \prod_{d=1}^D \zeta^{(d)\frac{\alpha}{D}-1} \int_{\tau} \left(\prod_{d=1}^D \prod_{m=1}^M \prod_{j_m=1}^{p_m} (\tau \zeta^{(d)} \sigma_m^2)^{-\frac{q_m}{2}} \right. \\ &\quad \cdot \exp \left\{ -\frac{1}{2} \left(\boldsymbol{\beta}_{m,j_m}^{(d)} - \boldsymbol{\gamma}_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m} \right)' (\tau \zeta^{(d)} \sigma_m^2)^{-1} \left(\boldsymbol{\beta}_{m,j_m}^{(d)} - \boldsymbol{\gamma}_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m} \right) \right\} \\ &\quad \cdot \left(\tau \zeta^{(d)} \mathbf{w}_{m,j_m}^{(d)} \right)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \frac{\gamma_{m,j_m}^{(d)2}}{\tau \zeta^{(d)} \mathbf{w}_{m,j_m}^{(d)}} \right\} \tau^{a_{\tau}-1} e^{-b_{\tau}\tau} d\tau \end{aligned}$$

$$\begin{aligned} &\propto \prod_{d=1}^D \zeta^{(d)\frac{\alpha}{D}-1} \int_{\tau} \left(\prod_{d=1}^D (\tau \zeta^{(d)})^{-\frac{\sum_{m=1}^M p_m(q_m+1)}{2}} \right. \\ &\cdot \exp \left\{ -\frac{1}{2\tau \zeta^{(d)}} \sum_{m=1}^M \sum_{j_m=1}^{p_m} \left(\frac{1}{\sigma_m^2} (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})' (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m}) + \frac{1}{w_{m,j_m}^{(d)}} \gamma_{m,j_m}^{(d)2} \right) \right\} \\ &\cdot \tau^{a_{\tau}-1} e^{-b_{\tau}\tau} d\tau \end{aligned}$$

Let us define $C_d = \sum_{m=1}^M \sum_{j_m=1}^{p_m} \left(\frac{1}{\sigma_m^2} (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})' (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m}) + \frac{1}{w_{m,j_m}^{(d)}} \gamma_{m,j_m}^{(d)2} \right)$,
 $I_0 = \sum_{m=1}^M p_m(q_m + 1) = M \prod_{m=1}^M p_m + \sum_{m=1}^M p_m$

$$p(\zeta | B, \gamma, \mathbf{w}) \propto \prod_{d=1}^D \zeta^{(d)\frac{\alpha}{D}-\frac{I_0}{2}-1} \int_{\tau} \tau^{a_{\tau}-\frac{DI_0}{2}-1} \exp \left\{ -b_{\tau}\tau - \frac{\sum_{d=1}^D C_d}{2\tau \zeta^{(d)}} \right\} d\tau$$

By definition, $\sum_{d=1}^D \zeta^{(d)} = 1$ which allows us to write $\sum_{d=1}^D (b_{\tau}\tau \zeta^{(d)}) = b_{\tau}\tau \sum_{d=1}^D \zeta^{(d)} = b_{\tau}\tau$ (Billio et al., 2023), moreover, by letting $a_{\tau} = \alpha$ (Guhaniyogi et al. (2017))

$$p(\zeta | B, \gamma, \mathbf{w}) \propto \int_{\tau} \left(\prod_{d=1}^D \zeta^{(d)\frac{\alpha}{D}-\frac{I_0}{2}-1} \right) \tau^{(\alpha-\frac{DI_0}{2})-1} \exp \left\{ -\frac{1}{2} \sum_{d=1}^D \left(\frac{C_d}{\tau \zeta^{(d)}} + 2b_{\tau}\tau \zeta^{(d)} \right) \right\} d\tau \quad (\text{B.4})$$

We recognize that equation (B.4) is the kernel of a Generalized Inverse Gaussian for $\phi^{(d)} = \tau \zeta^{(d)}$

$$\phi^{(d)} \sim \text{GiG} \left(\frac{\alpha}{D} - \frac{I_0}{2}, 2b_{\tau}, C_d \right)$$

We then obtain $\zeta^{(d)}$ by renormalizing: $\zeta^{(d)} = \phi^{(d)} / \sum_{l=1}^D \phi^{(l)}$ (Billio et al., 2023).

The full conditional for τ can be derived as follows

$$\begin{aligned} p(\tau | B, \gamma, \mathbf{w}, \boldsymbol{\zeta}) &\propto \pi(\tau) p(B, \gamma | \zeta^{(d)}, \tau, \mathbf{w}) = \pi(\tau) p(B | \gamma, \mathbf{w}, \zeta^{(d)}, \tau) p(\gamma | \mathbf{w}, \zeta^{(d)}, \tau) \\ &= \tau^{a_{\tau}-1} e^{-b_{\tau}\tau} \prod_{d=1}^D \prod_{m=1}^M \prod_{j_m=1}^{p_m} (\tau \zeta^{(d)} \sigma_m^2)^{-\frac{q_m}{2}} \\ &\cdot \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})' (\tau \zeta^{(d)} \sigma_m^2)^{-1} (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m}) \right\} \\ &\cdot (\tau \zeta^{(d)} w_{m,j_m}^{(d)})^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \frac{\gamma_{m,j_m}^{(d)2}}{\tau \zeta^{(d)} w_{m,j_m}^{(d)}} \right\} \\ &\propto \tau^{a_{\tau}-\frac{DI_0}{2}-1} e^{-b_{\tau}\tau} \\ &\cdot \prod_{d=1}^D \exp \left\{ -\frac{1}{2\tau \zeta^{(d)}} \sum_{m=1}^M \sum_{j_m=1}^{p_m} \left(\frac{1}{\sigma_m^2} (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})' (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m}) + \frac{1}{w_{m,j_m}^{(d)}} \gamma_{m,j_m}^{(d)2} \right) \right\} \\ &= \tau^{a_{\tau}-\frac{DI_0}{2}-1} \exp \left\{ -\frac{1}{2} \left(\sum_{d=1}^D \frac{C_d}{\tau \zeta^{(d)}} + 2b_{\tau}\tau \right) \right\} \end{aligned}$$

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

$$p(\tau | B, \gamma, \mathbf{w}, \boldsymbol{\zeta}) \propto \text{GiG} \left(a_{\tau} - \frac{DI_0}{2}, 2b_{\tau}, \sum_{d=1}^D \frac{C_d}{\zeta^{(d)}} \right)$$

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

Notice that by the construction of the prior distributions, $\gamma_{m,j_m}^{(d)}$ follows a double exponential distribution given $\lambda_m^{(d)}, \tau, \zeta^{(d)}$

$$\gamma_{m,j_m}^{(d)} \sim \text{DE} \left(0, \frac{\sqrt{\tau \zeta^{(d)}}}{\lambda_m^{(d)}} \right)$$

The full conditional of $\lambda_m^{(d)}$ can be written as

$$\begin{aligned} p \left(\lambda_m^{(d)} | \gamma_{m,j_m}^{(d)}, \tau, \zeta^{(d)} \right) &\propto \pi(\lambda_m^{(d)}) p \left(\gamma_{m,j_m}^{(d)} | \lambda_m^{(d)}, \tau, \zeta^{(d)} \right) \\ &\propto \left(\tau \zeta^{(d)} \right)^{-\frac{p_m}{2}} \left(\lambda_m^{(d)} \right)^{a_\lambda + p_m - 1} \exp \left\{ - \left(\frac{\sum_{j_m=1}^{p_m} |\gamma_{m,j_m}^{(d)}|}{\sqrt{\tau \zeta^{(d)}}} + b_\lambda \right) \lambda_m^{(d)} \right\} \\ &\propto \mathcal{G}a \left(a_\lambda + p_m, \frac{\sum_{j_m=1}^{p_m} |\gamma_{m,j_m}^{(d)}|}{\sqrt{\tau \zeta^{(d)}}} + b_\lambda \right) \end{aligned}$$

The full conditional for $w_{m,j_m}^{(d)}$ is

$$\begin{aligned} p \left(w_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)} \right) &\propto \pi \left(w_{m,j_m}^{(d)} \right) p \left(\gamma_{m,j_m}^{(d)} | \lambda_m^{(d)}, \tau, \zeta^{(d)}, w_{m,j_m}^{(d)} \right) \\ &\propto w_{m,j_m}^{(d) \frac{1}{2} - 1} \exp \left\{ - \frac{1}{2} \left(\lambda_m^{(d)2} w_{m,j_m}^{(d)} + \frac{\gamma_{m,j_m}^{(d)2}}{\tau \zeta^{(d)} w_{m,j_m}^{(d)}} \right) \right\} \\ &\propto \text{GiG} \left(\frac{1}{2}, \lambda_m^{(d)2}, \frac{\gamma_{m,j_m}^{(d)2}}{\tau \zeta^{(d)}} \right) \end{aligned}$$

Block 3: Sampling $\beta_{m,j_m}^{(d)}, \gamma_{m,j_m}^{(d)}, \sigma_m^2, \mu, \sigma^2$ from $p \left(\beta_{m,j_m}^{(d)}, \gamma_{m,j_m}^{(d)}, \sigma_m^2, \mu, \sigma^2 | \mathbf{y}, B_1, \dots, B_T \right)$

We derive the full conditional of $\beta_{m,j_m}^{(d)}$ in a way such that it only depends on observed data by integrating out $\gamma_{m,j_m}^{(d)}$. The total number of $\beta_{m,j_m}^{(d)}$ we need to sample is $D \sum_{m=1}^M p_m$.

$$\begin{aligned} p \left(\beta_{m,j_m}^{(d)} | \mathbf{y}, B_1, \dots, B_T, \sigma_m^2 \right) &\propto p \left(\mathbf{y} | \beta_{m,j_m}^{(d)}, \sigma_m^2, B_1, \dots, B_T \right) \int_\gamma p \left(\beta_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)}, \sigma_m^2 \right) p \left(\gamma_{m,j_m}^{(d)} \right) d\gamma \\ &\propto \prod_{t \in \mathcal{T}} \exp \left\{ - \frac{1}{2} \frac{(y_t - \mu - \langle B, B_t \rangle)^2}{\sigma^2} \right\} \int_\gamma p \left(\beta_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)}, \sigma_m^2 \right) p \left(\gamma_{m,j_m}^{(d)} \right) d\gamma \end{aligned}$$

where $\mathcal{T} \subset \{1, \dots, T\}$ contains all the indexes of the observations such that the latent variable s_t takes a given value in $\{1, \dots, K\}$. Thanks to the result in Proposition 2, the terms at the exponent in the likelihood write as:

$$\begin{aligned} &\left(y_t - \mu - \beta_{m,j_m}^{(d)'} \Psi_{j_m,m,t}^{(d)} - R_{j_m,m,t}^{(d)} - R_t^{(d)} \right)^2 = \\ &\left(y_t - \mu - R_{j_m,m,t}^{(d)} - R_t^{(d)} \right)^2 + \beta_{m,j_m}^{(d)'} \Psi_{j_m,m,t}^{(d)} \Psi_{j_m,m,t}^{(d)'} \beta_{m,j_m}^{(d)} - 2 \beta_{m,j_m}^{(d)'} \Psi_{j_m,m,t}^{(d)} \left(y_t - \mu - R_{j_m,m,t}^{(d)} - R_t^{(d)} \right) \end{aligned}$$

Define $\tilde{y}_t = y_t - \mu - R_{j_m,m,t}^{(d)} - R_t^{(d)}$, then the likelihood can be written as

$$p \left(\mathbf{y} | \beta_{m,j_m}^{(d)}, \sigma_m^2, B_1, \dots, B_T \right) \propto \exp \left\{ - \frac{1}{2\sigma^2} \sum_{t \in \mathcal{T}} \left(\beta_{m,j_m}^{(d)'} \Psi_{j_m,m,t}^{(d)} \Psi_{j_m,m,t}^{(d)'} \beta_{m,j_m}^{(d)} - 2 \beta_{m,j_m}^{(d)'} \Psi_{j_m,m,t}^{(d)} \tilde{y}_t \right) \right\}$$

For the integration part, given $p(\boldsymbol{\beta}_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)})$ and $p(\gamma_{m,j_m}^{(d)})$ are normal then the marginal distribution is normal with mean $\mathbb{E}(\boldsymbol{\beta}_{m,j_m}^{(d)}) = \mathbb{E}(\mathbb{E}(\boldsymbol{\beta}_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)})) = \mathbf{0}$, and variance $\mathbb{V}(\boldsymbol{\beta}_{m,j_m}^{(d)}) = \mathbb{V}(\mathbb{E}(\boldsymbol{\beta}_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)})) + \mathbb{E}(\mathbb{V}(\boldsymbol{\beta}_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)})) = (\tau\zeta^{(d)}w_{m,j_m}^{(d)} + \tau\zeta^{(d)}\sigma_m^2) I_{q_m}$.

Let $\xi = \tau\zeta^{(d)}(w_{m,j_m}^{(d)} + \sigma_m^2)$, then the full conditional of $\boldsymbol{\beta}_{m,j_m}^{(d)}$ can be written as

$$\begin{aligned} & p(\boldsymbol{\beta}_{m,j_m}^{(d)} | \mathbf{y}, B_1, \dots, B_T, \sigma_m^2) \\ & \propto \exp \left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_{m,j_m}^{(d)'} \left(\sum_{t \in \mathcal{T}} \frac{\Psi_{j_m,m,t}^{(d)} \Psi_{j_m,m,t}^{(d)'}}{\sigma^2} + \frac{1}{\xi} I_{q_m} \right) \boldsymbol{\beta}_{m,j_m}^{(d)} - 2\boldsymbol{\beta}_{m,j_m}^{(d)'} \sum_{t \in \mathcal{T}} \frac{\Psi_{j_m,m,t}^{(d)} \tilde{y}_t}{\sigma^2} \right] \right\} \\ & \propto \mathcal{N} \left(\left(\sum_{t \in \mathcal{T}} \frac{\Psi_{j_m,m,t}^{(d)} \Psi_{j_m,m,t}^{(d)'}}{\sigma^2} + \frac{1}{\xi} I_{q_m} \right)^{-1} \sum_{t \in \mathcal{T}} \frac{\Psi_{j_m,m,t}^{(d)} \tilde{y}_t}{\sigma^2}, \left(\sum_{t \in \mathcal{T}} \frac{\Psi_{j_m,m,t}^{(d)} \Psi_{j_m,m,t}^{(d)'}}{\sigma^2} + \frac{1}{\xi} I_{q_m} \right)^{-1} \right) \end{aligned}$$

The full conditional for $\gamma_{m,j_m}^{(d)}$ given $\boldsymbol{\beta}_{m,j_m}^{(d)}$ can be written as

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

The full conditional for σ_m^2 can be written as

$$\begin{aligned} & p(\sigma_m^2 | \boldsymbol{\beta}_{m,j_m}^{(d)}) \propto \prod_{d=1}^D \prod_{j_m=1}^{p_m} p(\boldsymbol{\beta}_{m,j_m}^{(d)} | \sigma_m^2) \pi(\sigma_m^2) \\ & = (\sigma_m^2)^{a_{\sigma_m} - D \frac{p_m q_m}{2} - 1} \exp \left\{ -\frac{1}{2} \left(\frac{\sum_{d=1}^D \sum_{j_m=1}^{p_m} (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})' (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})}{\tau\zeta^{(d)} \sigma_m^2} + 2b_{\sigma_m} \sigma_m^2 \right) \right\} \\ & \propto \text{GiG} \left(a_{\sigma_m} - D \frac{p_m q_m}{2}, 2b_{\sigma_m}, \frac{\sum_{d=1}^D \sum_{j_m=1}^{p_m} (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})' (\boldsymbol{\beta}_{m,j_m}^{(d)} - \gamma_{m,j_m}^{(d)} \boldsymbol{\nu}_{q_m})}{\tau\zeta^{(d)}} \right) \end{aligned}$$

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

$$\begin{aligned} & p(\sigma^2 | \mathbf{y}) \propto p(\mathbf{y} | \sigma^2) \pi(\sigma^2) \\ & \propto (\sigma^2)^{-(a_\sigma + \frac{T}{2}) - 1} \exp \left\{ -\frac{1}{\sigma^2} \left(\frac{1}{2} \sum_{t=1}^T (y_t - \langle B, X_t \rangle - \mu)^2 + b_\sigma \right) \right\} \propto \mathcal{IG}(a_\sigma^*, b_\sigma^*) \end{aligned}$$

where $a_\sigma^* = a_\sigma + \frac{T}{2}$ and $b_\sigma^* = \frac{1}{2} \sum_{t=1}^T (y_t - \langle B, X_t \rangle - \mu)^2 + b_\sigma$

Let $\mu^* = \sum_{t=1}^T (y_t - \langle B, X_t \rangle) \sigma_\mu^{*2}$ and $\sigma_\mu^{*2} = \left(\frac{T}{\sigma^2} + \frac{1}{\sigma_\mu^2} \right)^{-1}$, the full conditional of μ is:

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

References

- Agudze, K. M., Billio, M., Casarin, R., and Ravazzolo, F. (2022). Markov switching panel with endogenous synchronization effects. *Journal of Econometrics*, 230(2):281–298.
- Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119.
- Bandi, F. M. and Perron, B. (2006). Long memory and the relation between implied and realized volatility. *Journal of Financial Econometrics*, 4(4):636–670.
- Bianchi, D., Billio, M., Casarin, R., and Guidolin, M. (2019). Modeling systemic risk with Markov switching graphical sur models. *Journal of Econometrics*, 210(1):58 – 74.
- Billio, M., Casarin, R., and Iacopini, M. (2024). Bayesian Markov-switching tensor regression for time-varying networks. *Journal of the American Statistical Association (Theory & Methods)*, 119(545):109–121.
- Billio, M., Casarin, R., Iacopini, M., and Kaufmann, S. (2023). Bayesian dynamic tensor regression. *Journal of Business & Economic Statistics*, 41(2):429–439.
- Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H. (2016). Interactions between Eurozone and US Booms and Busts: A Bayesian Panel Markov-switching VAR model. *Journal of Applied Econometrics*, 31(1):1352–1370.
- Caffo, B. S., Crainiceanu, C. M., Verduzco, G., Joel, S., Mostofsky, S. H., Bassett, S. S., and Pekar, J. J. (2010). Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer’s disease risk. *NeuroImage*, 51(3):1140–1149.
- Casarin, R., Foroni, C., Marcellino, M., and Ravazzolo, F. (2018a). Economic uncertainty through the lenses of a mixed-frequency Bayesian panel Markov switching model. *Annals of Applied Statistics*, 12(4):2559 – 2568.
- Casarin, R., Sartore, D., and Tronzano, M. (2018b). A Bayesian Markov-switching correlation model for contagion analysis on exchange rate markets. *Journal of Business & Economic Statistics*, 36(1):101–114.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.
- Fernandes, M., Medeiros, M. C., and Scharth, M. (2014). Modeling and predicting the cboe market volatility index. *Journal of Banking & Finance*, 40:1–10.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.
- Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The midas touch: Mixed data sampling regression models. *Discussion Paper UCLA and UNC*.
- Golub, G. H. and Loan, C. F. V. (1989). *Computations, Matrix*. Johns Hopkins University Press.
- Guha, S. and Rodriguez, A. (2021). Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association*, 116(534):581–593.
- Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1):2733–2763.
- Gupta, A. K. and Nagar, D. K. (1999). *Matrix Variate Distributions*. Chapman & Hall, CRC Press.
- Hastie, T. and Tibshirani, R. (2000). Bayesian backfitting. *Statistical Science*, 15(3):196–213.

- Kaufmann, S. (2015). K-state Switching Models with Time-varying Transition Distributions: Does Loan Growth Signal Stronger Effects of Variables on Inflation? *Journal of Econometrics*, 187(1):82–94.
- Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.
- Kossaifi, J., Lipton, Z. C., Kolbeinsson, A., Khanna, A., Furlanello, T., and Anandkumar, A. (2020). Tensor regression networks. *The Journal of Machine Learning Research*, 21(1):4862–4882.
- Levine, R. A. and Casella, G. (2006). Optimizing random scan Gibbs samplers. *Journal of Multivariate Analysis*, 97(10):2071–2100.
- Newman, M. (2018). *Networks*. Oxford University Press.
- Papadogeorgou, G., Zhang, Z., and Dunson, D. B. (2021). Soft tensor regression. *J. Mach. Learn. Res.*, 22:219–1.
- Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.
- Spencer, D., Guhaniyogi, R., Shinohara, R., and Prado, R. (2022). Bayesian tensor regression using the Tucker decomposition for sparse spatial modeling. *arXiv preprint arXiv:2203.04733*.
- Wang, K. and Xu, Y. (2024). Bayesian tensor-on-tensor regression with efficient computation. *Statistics and its Interface*, 17(2):199.
- Xiao, J. and Wang, Y. (2022). Good oil volatility, bad oil volatility, and stock return predictability. *International Review of Economics & Finance*, 80:953–966.
- Xiao, J., Wang, Y., and Wen, D. (2023). The predictive effect of risk aversion on oil returns under different market conditions. *Energy Economics*, 126:106969.
- Yu, R., Li, G., and Liu, Y. (2018). Tensor regression meets Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 482–490. PMLR.
- Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 373–381, New York, New York, USA. PMLR.
- Zhang, Z., Allen, G. I., Zhu, H., and Dunson, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. *Neuroimage*, 197:330–343.
- Zhao, Q., Zhou, G., Zhang, L., and Cichocki, A. (2014). Tensor-variate Gaussian processes regression and its application to video surveillance. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1265–1269. IEEE.
- Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.
- Łatuszyński, K., Roberts, G. O., and Rosenthal, J. S. (2013). Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability*, 23(1):66 – 98.

Markov Switching Multiple-equation Tensor Regressions

Online supplementary material

This supplement consists of two Appendices C and D. Appendix C contains simulation results, while Appendix D includes more details on the empirical applications.

C Simulation results

C.1 Simple tensor regression

For the simple tensor regression, we designed four different experimental settings. In each setting, the dimension of the true coefficients is 20×20 , and the true rank value and sparsity level are different following the scenarios one can find in real-data applications from many fields, including neuroscience, image processing and network analysis.

- (\mathcal{S}_1) In the first setting, the true coefficient is a diagonal matrix representing a scenario of full rank and a high level of sparsity. It can be interpreted as the adjacency matrix of a graph without edges between distinct nodes (empty graph) configuration, where edges between a node and itself (self-loops) are allowed. See Newman (2018) for an introduction to networks.
- (\mathcal{S}_2) In the second setting, the cross pattern in the coefficient values returns a symmetric persymmetric matrix (Golub and Loan, 1989) with a lower level of sparsity and a lower true rank. It can be interpreted as the adjacency matrix of a graph with a hub, a node with a number of edges that greatly exceeds the average.
- (\mathcal{S}_3) The third setting is similar to the second and differs only in the configuration of the coefficient matrix, which contains a circular pattern. It can be interpreted as the adjacency matrix of a graph where the nodes are either disconnected from the other nodes or connected, directly or indirectly, to all the other nodes (connected component).
- (\mathcal{S}_4) In the last setting, the true coefficient matrix is symmetric, not persymmetric. It can be interpreted as the adjacency matrix of a core-periphery graph. A core-periphery structure is a common feature in social networks, where one expects to find a core of nodes with a high degree of connections surrounded by a periphery of nodes with a lower degree of connections.

Since the coefficients are binary in the above settings, we also consider more challenging settings in which Gaussian noise is added to the coefficient values to remove exact zeros. These real-valued settings are denoted by $\tilde{\mathcal{S}}_1$ to $\tilde{\mathcal{S}}_4$. For each setting, two types of data generating processes (DGP) are examined: i) with covariates randomly drawn *i.i.d* from a standard normal distribution; ii) with covariates following an AR(1) process. The number of observations generated from each DGP is 400. We check model performances and sampling efficiency under three rank orders: $D = \{3, 5, 7\}$.

We ran the proposed Gibbs sampler for 10,000 iterations. Figure C.1 shows the simulation results for the four settings with exact zeros in the true coefficients. Further illustrations of the simulation results for each setting are given Figure C.5. The panels show the results with AR(1) setting and for the noisy coefficients setting, respectively. In all scenarios, the algorithm can recover the true structure of the coefficients. All the ranks applied in the algorithm led to good estimations, and higher ranks, such as 5 or 7, do not offer significant improvements.

C.2 MS tensor regression

For the MSMETR, we consider two regimes with a coefficient matrix of size 12×12 . The sample size is 800, and two different experimental settings are considered.

- (\mathcal{S}_1^{MS}) In the first setting, the true coefficients are an anti-diagonal and diagonal matrix. They represent the setting where the true coefficients of the two regimes have similar levels of sparsity and ranks.
- (\mathcal{S}_2^{MS}) In the second setting, we replaced the anti-diagonal matrix with a cross matrix. They represent the setting where the two regimes' true coefficients vary in sparsity and rank.

Due to regime identification and label-switching issues (Frühwirth-Schnatter (2006)), an identification constraint disentangles the two regimes. We study the effectiveness of alternative identification

Setting	IID covariates			AR(1) covariates		
	$D = 3$	$D = 5$	$D = 7$	$D = 3$	$D = 5$	$D = 7$
(a) Simple regression with binary coefficients						
\mathcal{S}_1	0.0149	0.0251	0.0313	0.0144	0.0230	0.0431
\mathcal{S}_2	0.0498	0.0735	0.0801	0.0703	0.0822	0.1061
\mathcal{S}_3	0.0321	0.0570	0.0890	0.0666	0.0491	0.0830
\mathcal{S}_4	0.0255	0.0312	0.0550	0.0349	0.0467	0.0514
(b) Simple regression with real-valued coefficients						
$\tilde{\mathcal{S}}_1$	0.0290	0.0432	0.0508	0.0320	0.0421	0.0460
$\tilde{\mathcal{S}}_2$	0.0612	0.1193	0.1193	0.1222	0.0865	0.0920
$\tilde{\mathcal{S}}_3$	0.0709	0.0879	0.0809	0.0726	0.0751	0.0714
$\tilde{\mathcal{S}}_4$	0.0433	0.0622	0.0774	0.0538	0.0537	0.0738
(c) Markov-switching regression						
\mathcal{S}_1^{MS}	0.0059	0.0069	0.0087	0.0079	0.0063	0.0065
\mathcal{S}_2^{MS}	0.0075	0.0055	0.0121	0.0062	0.0072	0.0078

Table C.1: Mean square errors for the coefficients of the simple and switching regressions (panels a and b) in different settings (rows), data generating process and PARAFAC rank (columns).

Setting \mathcal{S}_1^{MS} (anti-diag / diag)					
	ACF(1)	ACF(5)	ACF(10)	MSE(10)	MSE(100)
Parameters (θ)	0.4085 (0.3145)	0.3279 (0.2328)	0.3158 (0.0980)	0.0559	0.0083
States(s_t)	0.5624 (0.5448)	0.5437 (0.3878)	0.5333 (0.1942)	0.2725	0.0113
Setting \mathcal{S}_2^{MS} (cross / diag)					
Parameters (θ)	0.5139 (0.4247)	0.4425 (0.2819)	0.4410 (0.1650)	0.1773	0.0106
States(s_t)	0.5294 (0.5153)	0.5166 (0.3649)	0.5077 (0.1831)	0.3013	0.0050

Table C.2: Results on convergence for the two different experimental settings. The second column reports the ACFs of the parameters and the hidden states before and after thinning. The results after thinning are reported in parentheses. The third column reports the MSE of the parameters and hidden states at the 10th and 100th Gibbs iteration.

strategies. In the first setting, we assume an identification constraint on the trace of the coefficients in the two regimes: $\text{tr}(B_1) < \text{tr}(B_2)$. In the second setting, use a constraint on the Frobenius norm $\|\cdot\|_F$ of the coefficients to identify the regimes, that is $\|B_1\|_F > \|B_2\|_F$.

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 entries. The number of Gibbs iterations in each experiment is 3000 with 1500 burn-in iterations.

The regression model and the MCMC algorithm provide reasonably accurate estimates of the true coefficient values, and the regimes are successfully identified in the two different experimental settings (see Figure C.2 for an illustration). The MSE of the actual outcome \mathbf{y} and estimated outcome $\hat{\mathbf{y}}$ is documented in the third column of Table C.2. The MSE decreases rapidly as the MCMC iterations increase, from 7.1119 at the 10th iteration to 1.1618 at the 100th iteration, and stabilizes around value 1.0 after convergence (see Figure C.4 in Appendix). The second column of Table C.2 shows the ACF of the parameters and the hidden states in two different experimental settings.

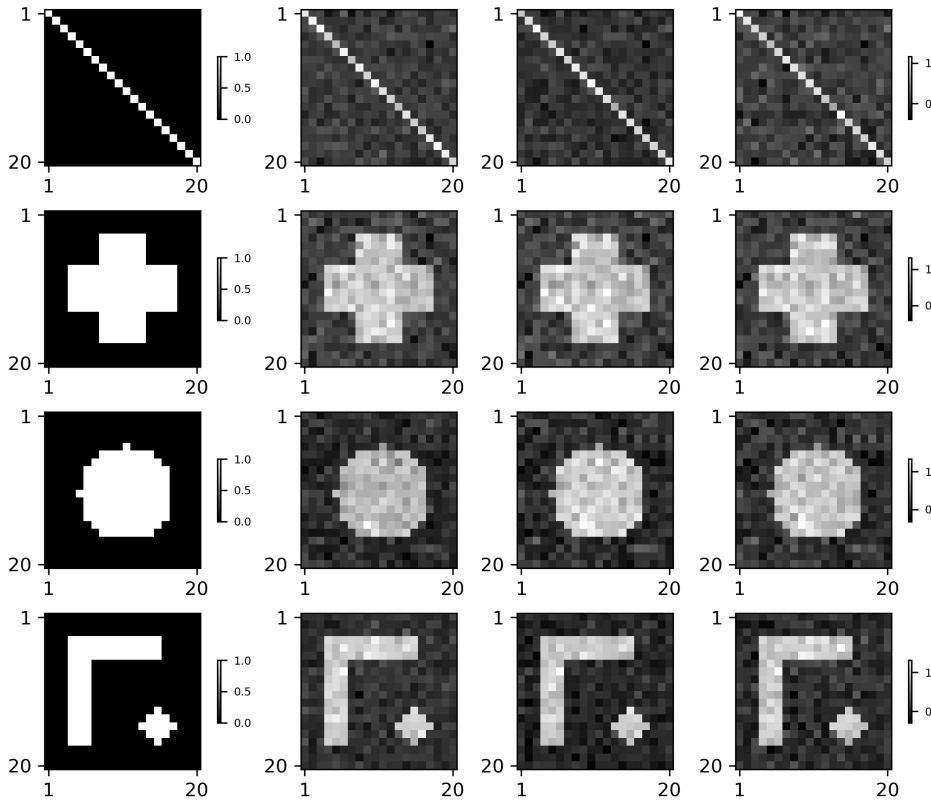


Figure C.1: First column contains the true coefficient value; columns from two to four contain the estimation results for different ranks $D = 3, 5, 7$. The four different experimental settings \mathcal{S}_j , $j = 1, \dots, 4$ are in the rows.

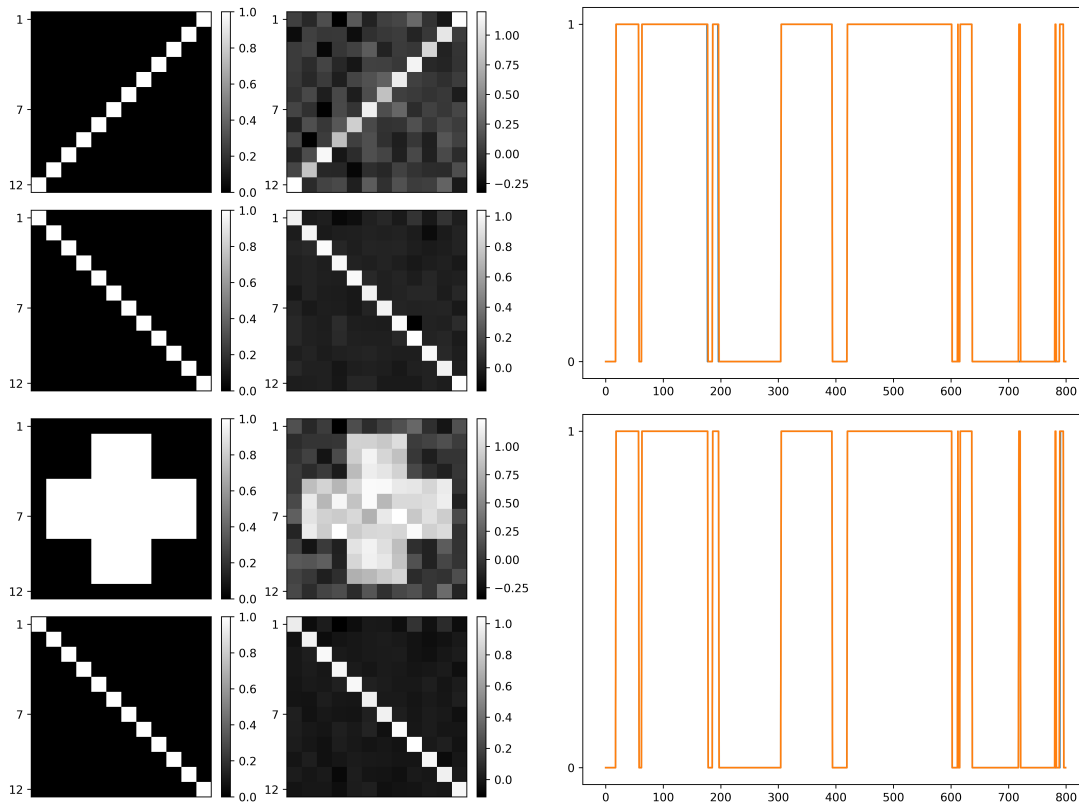


Figure C.2: Markov-switching model with Diagonal and Anti-diagonal coefficients (first row) and with Cross and Diagonal coefficients (second row). Left plots: true (left column) and estimated (right column) coefficients. Right plots: true (orange) and estimated (blue) hidden states.

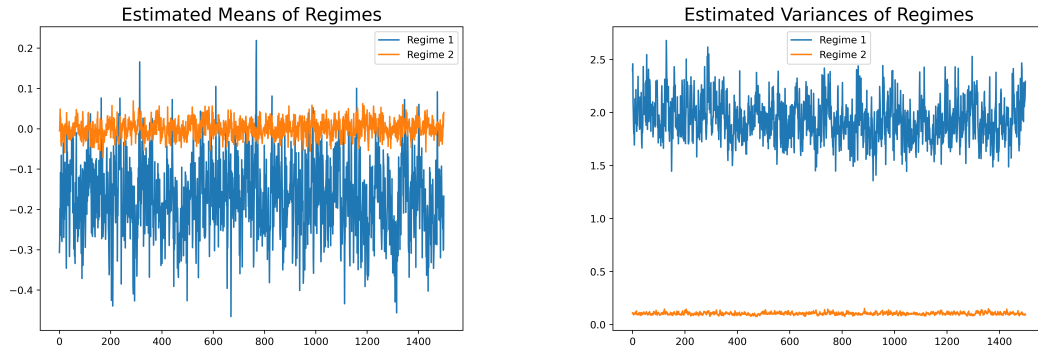


Figure C.3: Trace plots after removing the burn-in samples for regime-specific intercepts (left) and variances (right) for the \mathcal{S}_2^{MS} experimental setting. True values are $\mu_1 = \mu_2 = 0, \sigma_1^2 = 2, \sigma_2^2 = 0.1$.

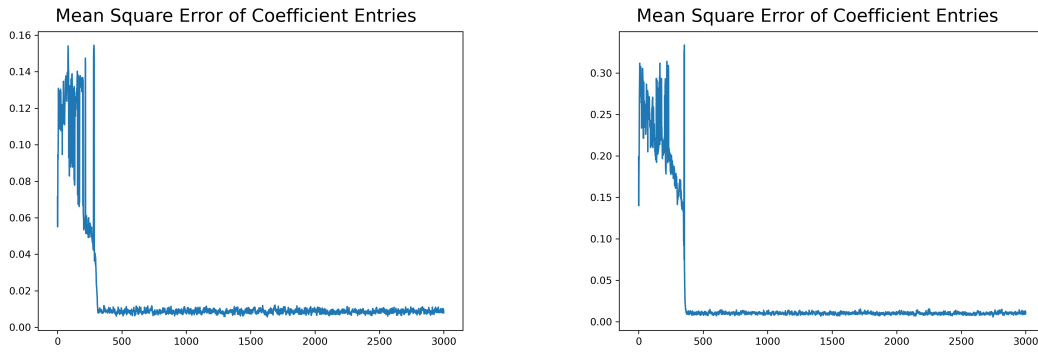


Figure C.4: MSE of the outputs for the \mathcal{S}_1^{MS} (left) and \mathcal{S}_2^{MS} (right) experimental setting.

C.3 Robustness checks

To test further the algorithm performance with respect to the choice of hyper-parameters, we carried out robustness checks by changing the prior mean and prior variance for σ_m^2, τ and $\lambda_m^{(d)}$. The simulation results for the new set of hyperparameters indicated in Table C.3 are given in Figure C.6. The weight inference results are robust in the choice of hyperparameter values.

Table C.3: Values of hyper-parameters chosen for benchmark and robustness check

	benchmark	robustness
α	1	1
a_σ	0.5	0.5
b_σ	$8.5\sqrt{C}$	$2\sqrt{C}$
a_τ	3	3
b_τ	$33.75/b_\sigma$	$33.75/b_\sigma$
a_λ	3	3
b_λ	$a_\lambda^{1/4}$	$a_\lambda^{1/2}$

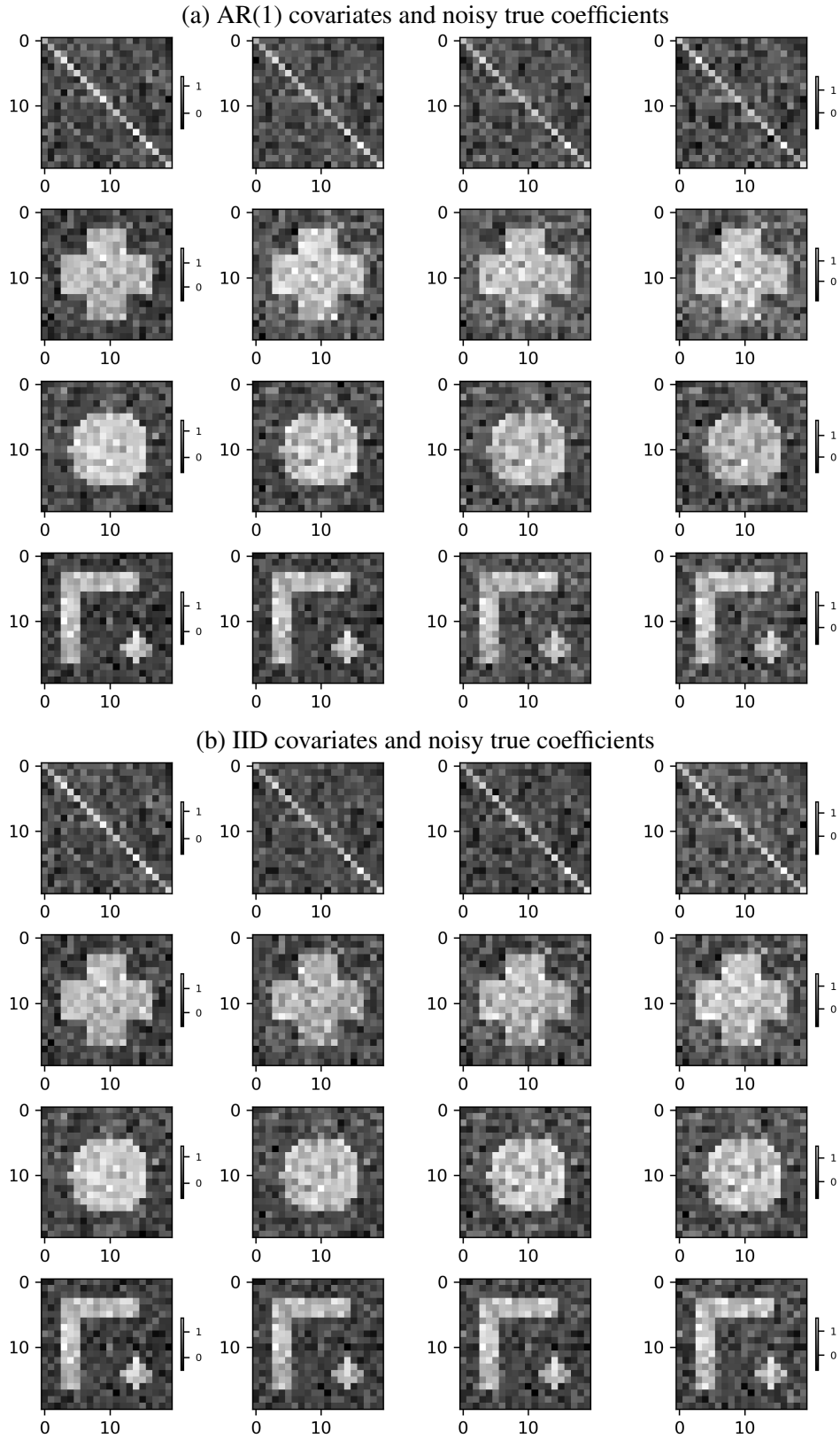


Figure C.5: In the first column, the true coefficient value, in the columns from two to four, the estimation results for different ranks $D = 3, 5, 7$. The four different experimental settings \mathcal{S}_j , $j = 1, \dots, 4$ are in the rows.

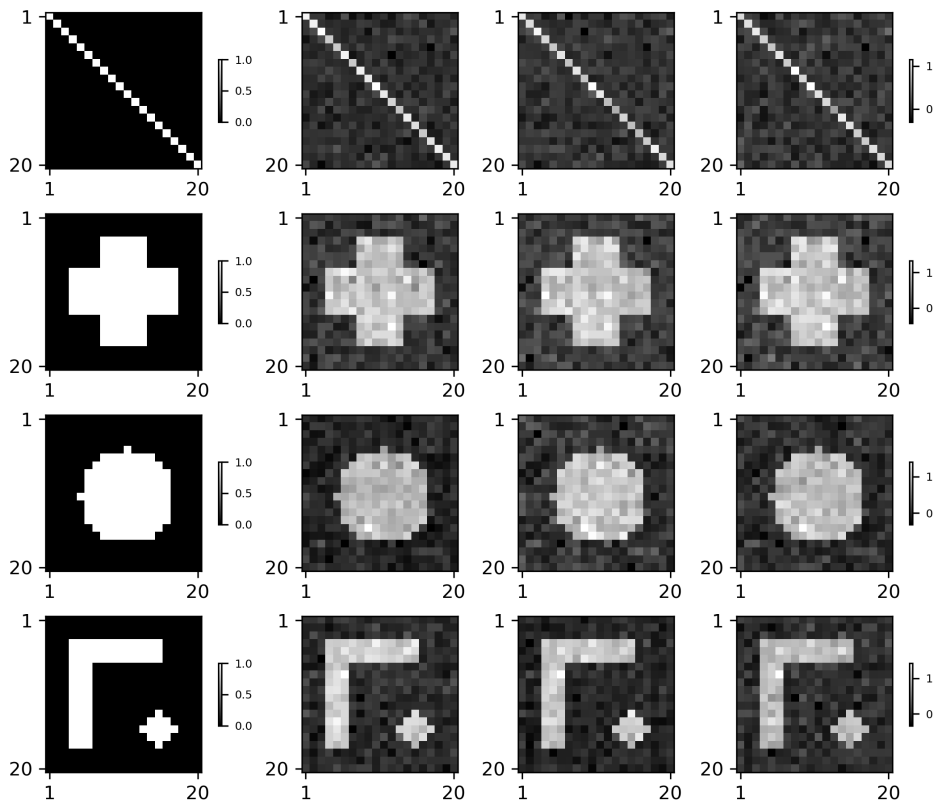


Figure C.6: Robustness check subject to a different set of hyper-parameters. In the first column, the true coefficient value, the estimation results for different ranks $D = 3, 5, 7$ in the columns from two to four. The four different experimental settings $\mathcal{S}_j, j = 1, \dots, 4$ are in the rows.

D Further Empirical Results

	In Sample		Out of Sample			
	MSE	MAE	1-day (month) ahead		5-day (month) ahead	
	MSE	MAE	MSE	MAE	MSE	MAE
<i>Application 1: VIX and OVX on macro indicators</i>						
Least Square	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
Tensor	0.3097	0.4324	0.2540	0.4232	0.3581	0.5182
MS Tensor	0.0907	0.2393	0.1409	0.3342	0.1379	0.3063
<i>Application 2: S&P 500 on oil prices (Aggregate Analysis)</i>						
Least Square	0.4418	0.5126	0.2394	0.4893	0.1986	0.4073
LASSO	0.4692	0.5264	0.2058	0.4537	0.1487	0.3548
Tensor	0.6073	0.6084	0.0445	0.2111	0.3442	0.4227
MS Tensor	0.3901	0.4794	0.5248	0.7244	0.2664	0.4522
<i>Application 2: Financial Sector on oil prices (Disaggregated S&P 500 analysis)</i>						
Least Square	0.5198	0.5530	0.3082	0.5551	0.1107	0.2631
LASSO	0.5532	0.5650	0.2684	0.5181	0.1111	0.2555
Tensor	0.6968	0.6204	0.0035	0.0594	0.0574	0.2025
MS Tensor	0.3370	0.4308	0.0586	0.2421	0.0878	0.2696
<i>Application 2: Energy Sector on oil prices (Disaggregated S&P 500 analysis)</i>						
Least Square	0.4587	0.5354	0.6606	0.8128	0.2879	0.4931
LASSO	0.4869	0.5516	0.1756	0.4191	0.2740	0.4545
Tensor	0.5379	0.5731	0.2414	0.4914	0.2788	0.4949
MS Tensor	0.3323	0.4362	0.0090	0.0949	0.3325	0.5153
<i>Application 2: Other sectors on oil prices (Disaggregated S&P 500 analysis)</i>						
Least Square	0.4389	0.5097	0.5919	0.7694	0.4054	0.5672
LASSO	0.4675	0.5254	0.3940	0.6277	0.2908	0.5071
Tensor	0.4643	0.5219	0.1072	0.3274	0.2935	0.4871
MS Tensor	0.3064	0.4105	0.9498	0.9746	0.7229	0.7488

Table D.1: In-sample fitting and out-of-sample forecasting performance. Results for the best-performing model are in boldface.

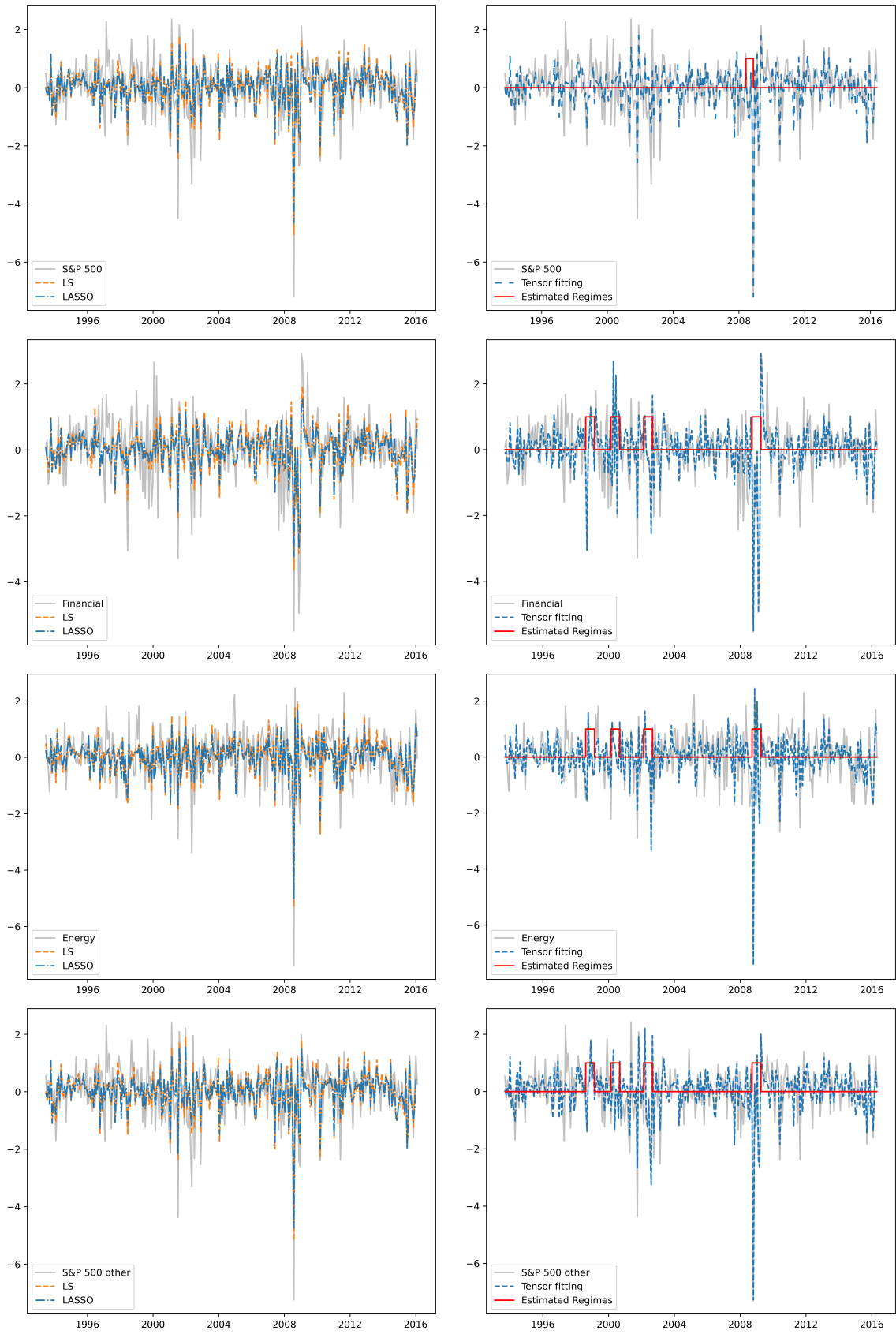


Figure D.1: Left: In-sample fitting results for Least Square (orange dashed line) and LASSO (blue dashed line). Right: Tensor Regression with Markov Switching (blue dashed line) and estimated hidden states (red solid line). True data is shown in solid silver line.

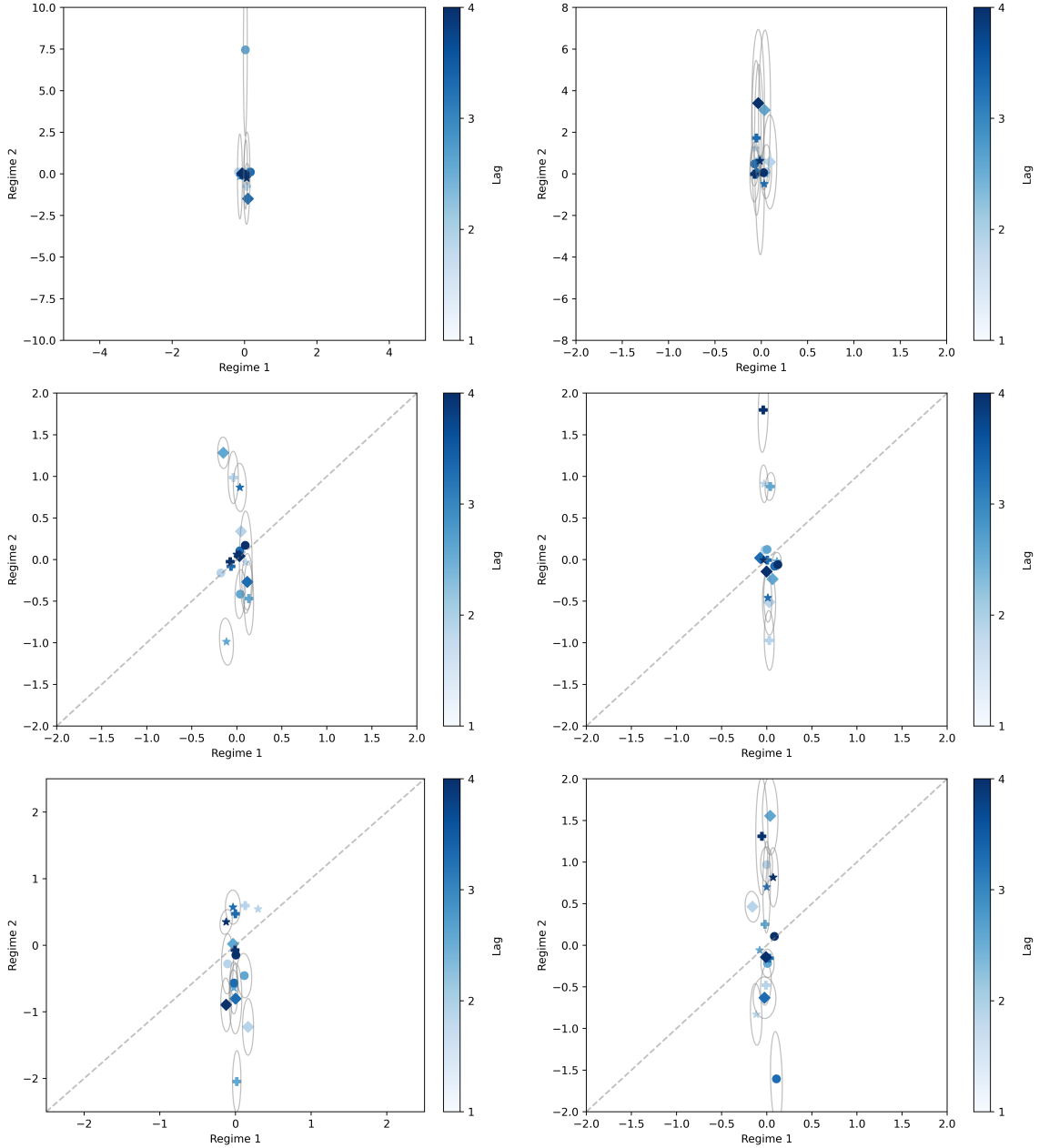


Figure D.2: The scatter plot shows the effects of Good Oil Volatility (left column) and Bad Oil Volatility (right column) on S&P 500 (first row), financial sector (second row) and energy sector (third row). Different symbols represent the weekly data sampled at different weeks, with \bullet : $t - (1 + 4(p - 1))/4$, \oplus : $t - 2(1 + 4(p - 1))/4$, \blacklozenge : $t - 3(1 + 4(p - 1))/4$ and \star : $t - 4(1 + 4(p - 1))/4$ for $p = \{1, 2, 3, 4\}$. Different shades of blue represent different lags, from order 1 (lighter) to order 4 (darker).