Bayesian Dynamic Tensor Regression

Monica Billio
Department of Economics, Ca' Foscari University of Venice
and
Roberto Casarin
Department of Economics, Ca' Foscari University of Venice
and
Matteo Iacopini
Department of Econometrics and Data Science, Vrije Universiteit Amsterdam
Tinbergen Institute
and
Sylvia Kaufmann
Study Center Gerzensee, Foundation of the Swiss National Bank

January 19, 2022

Abstract

High- and multi-dimensional array data are becoming increasingly available. They admit a natural representation as tensors and call for appropriate statistical tools. We propose a new linear autoregressive tensor process (ART) for tensor-valued data, that encompasses some well-known time series models as special cases. We study its properties and derive the associated impulse response function. We exploit the PARAFAC low-rank decomposition for providing a parsimonious parametrization and develop a Bayesian inference allowing for shrinking effects. We apply the ART model to time series of multilayer networks and study the propagation of shocks across nodes, layers and time.

Keywords: Bayesian inference; dynamic networks; forecasting; multidimensional autoregression; tensor models

1 Introduction

Many modern datasets in applied science have a complex and multidimensional structure which is naturally represented by multidimensional arrays, or tensors (e.g., Hackbusch, 2012). In statistics and machine learning, tensor algebra provides a fundamental background for effective modeling and efficient algorithm design in big data handling (e.g. Cichocki, 2014). The increasing availability of long temporal sequences of tensor-valued data, such as multidimensional tables (Balazsi et al., 2015), multidimensional panel data (Kapetanios et al., 2021), multilayer networks (Aldasoro and Alves, 2018), electroencephalogram (a.k.a. EEG, Li and Zhang, 2017), neuroimaging (Zhou et al., 2013) has put forward some limitations of the existing multivariate time series models. A naïve approach to model tensors ignores the intrinsic structure of the data and fits a multivariate regression on the vectorized tensor data. However, this might result in inefficient estimation and misleading results (Yuan and Zhang, 2016), thus making such representations unsuited for tensor-valued data.

Tensor modeling in statistics is in its infancy and most of the research in this field has focused on the analysis of cross-sectional data, as applied in neuroimaging (e.g., functional magnetic resonance image, a.k.a. fMRI, EEG) and signal processing, whereas the literature on tensors in time series analysis is scarce. Most often, a tensor-valued covariate is used to predict a scalar outcome (e.g., see Guhaniyogi et al., 2017; Xu et al., 2013; Zhou et al., 2013), and only a few papers analyze tensor-on-tensor regression models (e.g., see Lock, 2018). Estimation of tensor regressions requires parameter regularization or dimension reduction since the number of entries of the coefficient tensor is larger than the sample size.

In contrast to the existing literature, this article introduces dynamics in tensor regression models by defining a new framework for linear time series regression with tensor-valued response and covariates. We study the properties of the stochastic process, such as stationarity, and derive impulse response functions. Standard multivariate regression models are obtained as special cases. To address the dimensionality challenges of dynamic tensor models, we propose a low-rank representation of the coefficient tensor and impose parameter regularization based on the shrinkage prior distribution of Guhaniyogi et al.

(2017).

Guhaniyogi et al. (2017) design a predictive model in a cross-sectional setting to investigate the relationship between a scalar medical index and matrix-valued brain images. Instead, we propose a new framework for dynamic tensor-on-tensor regression, and use it to investigate multilayer international economic networks.

Recent papers on tensor regression exploit tensor-valued covariates to predict a scalar outcome in a generalized linear model (Xu et al., 2013; Zhou et al., 2013), whereas Li et al. (2018) use the Tucker decomposition to propose low-rank approximations to the coefficient tensor. On the other hand, the tensor-on-vector regression is an alternative approach used to assess the impact of a vector of factors on a tensor-valued observable. Rabusseau and Kadri (2016) consider a higher-order low-rank regression, which is a tensor-on-vector linear model with a low-rank constraint on the coefficient tensor. They propose an algorithm to obtain an approximate solution to the restricted least squares problem. In a related contribution, Guha and Rodriguez (2020) develop a Bayesian linear model for assessing the impact of vector covariates on matrix-valued MRIs for several patients. They adopt a symmetric parallel factor (PARAFAC) decomposition to identify the tensor nodes and cells related to each predictor. To study the impact of one or more external stimuli or predictors on the human brain, Guhaniyogi and Spencer (2021) have developed a regression framework with a tensor response and scalar covariates, coupled with a novel multiway stick-breaking shrinkage prior distribution on the coefficient tensor. The method has been extended by Spencer et al. (2020) to an additive mixed regression model with a tensor response, with region-specific random effects to capture the connectivity between the measurements on a set of pre-specified groups of brain voxels. In the presence of structured tensor-response variables, such as maps of neural connections in the brain, Guha and Guhaniyogi (2021) have proposed a Bayesian generalized linear model with a symmetric tensor response and scalar predictors. A brief review of the most recent contributions on tensor regression models is presented in Guhaniyogi (2020).

Another stream of the literature considers regression models with tensor-valued responses and covariates. Hoff (2015) employs the Tucker product to define a tensor-on-tensor regression, generalizing the standard bilinear to a multilinear model. Tensors have also been used in the analysis of large multivariate categorical response vectors (Zhou et al., 2015) and in high-dimensional classification problems (Yang and Dunson, 2016). Extending all these approaches, we consider a novel linear autoregressive model for real-valued tensor response and covariates, and we apply it in a time series framework to investigate dynamic multilayer networks.

We exploit the contracted product, an operator that generalizes the Cayley matrix multiplication to tensors (Behera et al., 2020; Ji and Wei, 2018; Wang et al., 2020), to introduce a new autoregressive tensor model (ART) which generalizes the existing tensor regression frameworks along two lines. First, the ART model introduces dynamics in linear tensor regression and provides the tools for analyzing shock propagation in multidimensional dynamical systems. Second, we allow for both tensor-valued outcomes and covariates, a more general framework encompassing existing tensor as well as multivariate linear models (e.g., vector autoregressions, or VARs). Taking advantage of the properties of the contracted product, we derive new results on tensor algebra and study the main properties of the ART process. Besides, we derive the impulse response function and the forecast error variance decomposition for making predictions and analyzing shock propagation in the system.

Besides handling multidimensional data, tensor regression models are usually characterized by a high dimensional parameter space, which calls for the use of dimension reduction or shrinkage estimation techniques. Li and Zhang (2017) define a tensor-response linear regression on a vector covariate for studying the relationship between brain activity and individual control variables, using cross-sectional data. They use the envelope method for estimation, which assumes that part of the response variables (a set of linear combinations of them) is irrelevant to the regression. Moreover, their optimization framework depends on tuning parameters (e.g., the envelope dimensions), the choice of which depends on the tensor dimensions and the signal-to-noise ratio (i.e., the degree of sparsity). Here, we propose to use a PARAFAC representation (Hackbusch, 2012) of the coefficient tensor to obtain a parsimonious parametrization of the ART.

Parameter regularization and sparse estimation in high-dimensional models can be achieved through alternative approaches, such as the Lasso (Zhou et al., 2013), the spike-and-slab (Guha and Rodriguez, 2020), and the envelope method (Li and Zhang, 2017). Alternative approaches induce element-wise sparsity or assume reduced-rank coefficient tensors. In neuroimaging, Sun and Li (2017) propose a regression framework for a tensor response and a vector predictor, where the coefficient tensor embeds both types of sparse structures. Raskutti et al. (2019) derive general risk bounds of the estimated coefficient in high-dimensional tensor regression problems with several regularizers, such as Lasso penalty and reduced-rank. Goldsmith et al. (2014) develop scalar-on-3D-image regression that includes a latent binary indicator to discriminate between image locations with predictive and non-predictive power. Here, we adopt the more flexible regularization approach based on the global-local shrinkage prior developed in Guhaniyogi et al. (2017). In particular, we impose this prior on the marginal vectors of the PARAFAC representation of the coefficient tensor and we show that, for rank-1 coefficient tensor, the conditional prior on the entries is a Meijer-G prior with heavier tails than the Normal distribution (e.g, see Zhang et al., 2020).

The literature on network data modeling has rapidly increased after the recent financial crisis, both in theoretical and empirical analyses. Dynamic tensor models are a natural framework for the analysis of multilayer network data in finance, biology, and sociology. An example of a time series of network data consists of a collection of yearly snapshots of interbank or international trade networks. However, despite dynamic models may be more adequate for studying network data collected over time, most statistical models for network data remained static so far (De Paula, 2017). Few attempts have been made to model time-varying networks (e.g., Anacleto and Queen, 2017; Hoff, 2015), and most of the existing approaches focus on providing a representation and a description of temporally evolving graphs (e.g., Holme and Saramäki, 2012; Kostakos, 2009). We contribute to this literature by providing an original study of time-varying economic and financial networks and show that our dynamic tensor model can be used successfully to carry out impulse response analysis in a multidimensional setting.

The remainder of the paper is organized as follows. Section 2 provides an introduction to tensor algebra and presents the new modeling framework. Section 3 discusses parametrization strategies and a Bayesian inference procedure. Section 4 provides an empirical application and Section 5 gives some concluding remarks. Further details and results are provided in the supplementary material.

2 A Dynamic Tensor Model

In this section, we present a dynamic tensor regression model and discuss some of its properties and special cases. We review some notions of multilinear algebra which will be used in this paper, and refer the reader to the supplement for novel results on tensor algebra and further details.

2.1 Tensor Calculus and Decompositions

The use of tensors is well established in physics and mechanics (e.g., Abraham et al., 2012; Aris, 2012), but few contributions have been made beyond these disciplines. For a general introduction to the algebraic properties of tensor spaces, see Hackbusch (2012). Noteworthy introductions to operations on tensors and tensor decompositions are Lee and Cichocki (2018) and Kolda and Bader (2009), respectively.

A N-order real-valued tensor is a N-dimensional array $\mathcal{X} = (\mathcal{X}_{i_1,\ldots,i_N}) \in \mathbb{R}^{I_1\times\ldots\times I_N}$ with entries $\mathcal{X}_{i_1,\ldots,i_N}$ with $i_n = 1,\ldots,I_n$ and $n = 1,\ldots,N$. The *order* is the number of dimensions (also called modes). Vectors and matrices are examples of 1- and 2-order tensors, respectively. In the rest of the paper we will use lower-case letters for scalars, lower-case bold letters for vectors, capital letters for matrices and calligraphic capital letters for tensors. We use the symbol ":" to indicate selection of all elements of a given mode of a tensor. The mode-k *fiber* is the vector obtained by fixing all but the k-th index of the tensor, i.e. the equivalent of rows and columns in a matrix. Tensor *slices* and their generalizations, are obtained by keeping fixed all but two or more dimensions of the tensor.

It can be shown that the set of N-order tensors $\mathbb{R}^{I_1\times\ldots\times I_N}$ endowed with the standard addition $\mathcal{A}+\mathcal{B} = (\mathcal{A}_{i_1,\ldots,i_N} + \mathcal{B}_{i_1,\ldots,i_N})$ and scalar multiplication $\alpha\mathcal{A} = (\alpha\mathcal{A}_{i_1,\ldots,i_N})$, with $\alpha \in \mathbb{R}$, is a vector space. We now introduce some operators on the set of real tensors, starting with the *contracted product*, which generalizes the matrix product to tensors. The contracted product between $\mathcal{X} \in \mathbb{R}^{I_1\times\ldots\times I_M}$ and $\mathcal{Y} \in \mathbb{R}^{J_1\times\ldots\times J_N}$ with $I_M = J_1$, is denoted by $\mathcal{X} \times_M \mathcal{Y}$ and yields a $(M+N-2)$-order tensor $\mathcal{Z} \in \mathbb{R}^{I_1 \times \ldots \times I_{M-1} \times J_1 \times \ldots \times J_{N-1}}$, with entries

$$\mathcal{Z}_{i_1,\ldots,i_{M-1},j_2,\ldots,j_N} = (\mathcal{X} \times_M \mathcal{Y})_{i_1,\ldots,i_{M-1},j_2,\ldots,j_N} = \sum_{i_M=1}^{I_M} \mathcal{X}_{i_1,\ldots,i_{M-1},i_M} \mathcal{Y}_{i_M,j_2,\ldots,j_N}.$$

When $\mathcal{Y} = \mathbf{y}$ is a vector, the contracted product is also called *mode-M product*. We define with $\mathcal{X} \bar{\times}_N \mathcal{Y}$ a sequence of contracted products between the $(K+N)$-order tensor $\mathcal{X} \in \mathbb{R}^{J_1 \times \ldots \times J_K \times I_1 \times \ldots \times I_N}$ and the $(N+M)$-order tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots \times I_N \times H_1 \times \ldots \times H_M}$. Entry-wise, it is defined as

$$\left(\mathcal{X} \bar{\times}_N \mathcal{Y}\right)_{j_1,\ldots,j_K,h_1,\ldots,h_M} = \sum_{i_1=1}^{I_1} \ldots \sum_{i_N=1}^{I_N} \mathcal{X}_{j_1,\ldots,j_K,i_1,\ldots,i_N} \mathcal{Y}_{i_1,\ldots,i_N,h_1,\ldots,h_M}.$$

Note that the contracted product is not commutative. The *outer product* $\circ$ between a M-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_M}$ and a N-order tensor $\mathcal{Y} \in \mathbb{R}^{J_1 \times \ldots \times J_N}$ is a $(M+N)$-order tensor $\mathcal{Z} \in \mathbb{R}^{I_1 \times \ldots \times I_M \times J_1 \times \ldots \times J_N}$ with entries $\mathcal{Z}_{i_1,\ldots,i_M,j_1,\ldots,j_N} = (\mathcal{X} \circ \mathcal{Y})_{i_1,\ldots,i_M,j_1,\ldots,j_N} = \mathcal{X}_{i_1,\ldots,i_M} \mathcal{Y}_{j_1,\ldots,j_N}$.

Tensor decompositions allow to represent a tensor as a function of lower dimensional variables, such as matrices of vectors, linked by suitable multidimensional operations. In this paper, we use the low-rank parallel factor (PARAFAC) decomposition, which allows to represent a N-order tensor in terms of a collection of vectors (called marginals). A N-order tensor is of rank 1 when it is the outer product of N vectors. Let R be the rank of the tensor $\mathcal{X}$, that is minimum number of rank-1 tensors whose linear combination yields $\mathcal{X}$. The PARAFAC(R) decomposition is rank-R decomposition which represents a N-order tensor $\mathcal{B}$ as a finite sum of R rank-1 tensors $\mathcal{B}_r$ defined by the outer products of N vectors (called marginals) $\boldsymbol{\beta}_j^{(r)} \in \mathbb{R}^{I_j}$

$$\mathcal{B} = \sum_{r=1}^{R} \mathcal{B}_r = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_N^{(r)}, \qquad \mathcal{B}_r = \boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_N^{(r)}. \tag{1}$$

The *mode-n matricization* (or unfolding), denoted by $\mathbf{X}_{(n)} = \mathrm{mat}_n(\mathcal{X})$, is the operation of transforming a N-dimensional array $\mathcal{X}$ into a matrix. It consists in re-arranging the mode-n fibers of the tensor to be the columns of the matrix $\mathbf{X}_{(n)}$, which has size $I_n \times I^*_{(-n)}$ with $I^*_{(-n)} = \prod_{i \neq n} I_i$. The mode-$n$ matricization of $\mathcal{X}$ maps the $(i_1, \ldots, i_N)$ element of $\mathcal{X}$ to the (i_n, j) element of $\mathbf{X}_{(n)}$, where $j = 1 + \sum_{m \neq n} (i_m - 1) \prod_{p \neq n}^{m-1} I_p$. For some numerical examples, see Kolda and Bader (2009) and Section S.1 in the supplement. The mode-1 unfolding is of interest for providing a visual representation of a tensor: for example, when $\mathcal{X}$ be a 3-order tensor, its mode-1 matricization $\mathbf{X}_{(1)}$ is a $I_1 \times I_2 I_3$ matrix obtained by horizontally stacking the mode-$(1,2)$ slices of the tensor. The *vectorization* operator stacks all the elements in direct lexicographic order, forming a vector of length $I^* = \prod_i I_i$. Other orderings are possible, as long as it is consistent across the calculations. The mode-n matricization can also be used to vectorize a tensor $\mathcal{X}$, by exploiting the relationship $\operatorname{vec}\left(\mathcal{X}\right) = \operatorname{vec}\left(\mathbf{X}_{(1)}\right)$, where $\operatorname{vec}\left(\mathbf{X}_{(1)}\right)$ stacks vertically into a vector the columns of the matrix $\mathbf{X}_{(1)}$. Many product operations have been defined for tensors (e.g. Lee and Cichocki, 2018), but here we constrain ourselves to the operators used in this work. For the ease of notation, we will use the multiple-index summation for indicating the sum over all the corresponding indices.

Remark 2.1. *Consider a N-order tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ with a PARAFAC(R) decomposition (with marginals $\boldsymbol{\beta}_j^{(r)}$), a $(N-1)$-order tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots \times I_{N-1}}$ and a vector $\mathbf{x} \in \mathbb{R}^{I_N}$. Then*

$$\mathcal{Y} = \mathcal{B} \times_N \mathbf{x} \iff \operatorname{vec}\left(\mathcal{Y}\right) = \mathbf{B}'_{(N)}\mathbf{x} \iff \operatorname{vec}\left(\mathcal{Y}\right)' = \mathbf{x}'\mathbf{B}_{(N)}$$

where $\mathbf{B}_{(N)} = \sum_{r=1}^{R} \boldsymbol{\beta}_N^{(r)} \operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_{N-1}^{(r)}\right)'$.

2.2 A General Dynamic Tensor Model

Let $\mathcal{Y}_t$ be a $(I_1 \times \ldots \times I_N)$-dimensional tensor of endogenous variables, $\mathcal{X}_t$ a $(J_1 \times \ldots \times J_M)$-dimensional tensor of covariates, and $S_y = \bigtimes_{j=1}^{N}\{1, \ldots, I_j\} \subset \mathbb{N}^N$ and $S_x = \bigtimes_{j=1}^{M}\{1, \ldots, J_j\} \subset \mathbb{N}^M$ sets of n-tuples of integers. We define the autoregressive tensor model of order p, ART(p), as the system of equations

$$\mathcal{Y}_{\mathbf{i},t} = \mathcal{A}_{\mathbf{i},0} + \sum_{j=1}^{p} \sum_{\mathbf{k} \in S_y} \mathcal{A}_{\mathbf{i},\mathbf{k},j} \mathcal{Y}_{\mathbf{k},t-j} + \sum_{\mathbf{m} \in S_x} \mathcal{B}_{\mathbf{i},\mathbf{m}} \mathcal{X}_{\mathbf{m},t} + \mathcal{E}_{\mathbf{i},t}, \quad \mathcal{E}_{\mathbf{i},t} \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\mathbf{i}}^2), \tag{2}$$

$t \in \mathbb{Z}$, with given initial conditions $\mathcal{Y}_{-p+1}, \ldots, \mathcal{Y}_0 \in \mathbb{R}^{I_1 \times \ldots \times I_N}$, where $\mathbf{i} = (i_1, \ldots, i_N) \in S_y$ and $\mathcal{Y}_{\mathbf{i},t}$ is the $\mathbf{i}$-th entry of $\mathcal{Y}_t$. The general model in Eq. (2) allows for measuring the effect of all the cells of $\mathcal{X}_t$ and of the lagged values of $\mathcal{Y}_t$ on each endogenous variable.

We give two equivalent compact representations of the multilinear system (2). The first one is used for studying the stability property of the process and is obtained through the contracted product that provides a natural setting for multilinear forms, decompositions and inversions. From (2) one gets

$$\mathcal{Y}_t = \mathcal{A}_0 + \sum_{j=1}^{p} \widetilde{\mathcal{A}}_j \bar{\times}_N \mathcal{Y}_{t-j} + \widetilde{\mathcal{B}} \bar{\times}_M \mathcal{X}_t + \mathcal{E}_t, \quad \mathcal{E}_t \overset{iid}{\sim} \mathcal{N}_{I_1,\ldots,I_N}(\mathcal{O}, \Sigma_1, \ldots, \Sigma_N), \tag{3}$$

$t \in \mathbb{Z}$, where $\bar{\times}_{a,b}$ is a shorthand notation for the contracted product $\times^{1\ldots a}_{a+1\ldots a+b}$ and $\bar{\times}_a$ is equivalent to $\bar{\times}_{a,0}$, $\widetilde{\mathcal{A}}_0$ is a N-order tensor of the same size as $\mathcal{Y}_t$, $\widetilde{\mathcal{A}}_j$, $j = 1, \ldots, p$, are $2N$-order tensors of size $(I_1 \times \ldots \times I_N \times I_1 \times \ldots \times I_N)$ and $\mathcal{B}$ is a $(N+M)$-order tensor of size $(I_1 \times \ldots \times I_N \times J_1 \times \ldots \times J_M)$. The error term $\mathcal{E}_t$ follows a N-order tensor normal distribution (Ohlson et al., 2013) with probability density function

$$f_{\mathcal{E}}(\mathcal{E}) = \frac{\exp\Big(-\frac{1}{2}(\mathcal{E} - \mathcal{M}) \bar{\times}_N \big(\circ_{j=1}^{N} \Sigma_j^{-1} \big) \bar{\times}_N (\mathcal{E} - \mathcal{M}) \Big)}{(2\pi)^{I^*/2} \prod_{j=1}^{N} \big|\Sigma_j\big|^{I^*_{-j}/2}}, \tag{4}$$

where $I^* = \prod_i I_i$ and $I^*_{-i} = \prod_{j \neq i} I_j$, $\mathcal{E}$ and $\mathcal{M}$ are N-order tensors of size $I_1 \times \ldots \times I_N$. Each covariance matrix $\Sigma_j \in \mathbb{R}^{I_j \times I_j}$, $j = 1, \ldots, N$, accounts for the dependence along the corresponding mode of $\mathcal{E}$.

The second representation of the ART(p) in Eq. (2) is used for developing inference. Let $\mathcal{K}_m$ be the $(I_1 \times \ldots \times I_N \times m)$-dimensional commutation tensor such that $\mathcal{K}_m^{\sigma} \bar{\times}_{N,0} \mathcal{K}_m = \mathbf{I}_m$, where $\mathcal{K}_m^{\sigma}$ is the tensor obtained by flipping the modes of $\mathcal{K}_m$. Define the $(I_1 \times \ldots \times I_N \times I^*)$-dimensional tensor $\mathcal{A}_j = \widetilde{\mathcal{A}}_j \bar{\times}_N \mathcal{K}_{I^*}$ and the $(I_1 \times \ldots \times I_N \times J^*)$-dimensional tensor $\mathcal{B} = \widetilde{\mathcal{B}} \bar{\times}_N \mathcal{K}_{J^*}$, with $J^* = \prod_j J_j$. We obtain $\mathcal{A}_j \times_{N+1} \mathrm{vec}\left(\mathcal{Y}_{t-j}\right) = \widetilde{\mathcal{A}}_j \bar{\times}_N \mathcal{Y}_{t-j}$ and the compact representation

$$\begin{aligned} &\mathcal{Y}_t = \mathcal{A}_0 + \sum_{j=1}^{p} \mathcal{A}_j \times_{N+1} \mathrm{vec}\left(\mathcal{Y}_{t-j}\right) + \mathcal{B} \times_{N+1} \mathrm{vec}\left(\mathcal{X}_t\right) + \mathcal{E}_t, \\ &\mathcal{E}_t \overset{iid}{\sim} \mathcal{N}_{I_1,\ldots,I_N}(\mathcal{O}, \Sigma_1, \ldots, \Sigma_N), \qquad t \in \mathbb{Z}. \end{aligned} \tag{5}$$

Let $\mathbb{T} = (\mathbb{R}^{I_1 \times \ldots \times I_N \times I_1 \times \ldots \times I_N}, \bar{\times}_N)$ be the space of $(I_1 \times \ldots \times I_N \times I_1 \times \ldots \times I_N)$-dimensional tensors endowed with the contracted product $\bar{\times}_N$. We define the identity tensor $\mathcal{I} \in \mathbb{T}$ to be the neutral element of $\bar{\times}_N$, that is the tensor whose entries are $\mathcal{I}_{i_1,\ldots,i_N,i_{N+1},\ldots,i_{2N}} = 1$ if $i_k = i_{k+N}$ for all $k = 1, \ldots, N$ and 0 otherwise. The inverse of a tensor $\mathcal{A} \in \mathbb{T}$ is the tensor

$\mathcal{A}^{-1} \in \mathbb{T}$ satisfying $\mathcal{A}^{-1} \bar{\times}_N \mathcal{A} = \mathcal{A} \bar{\times}_N \mathcal{A}^{-1} = \mathcal{I}$. A complex number $\lambda \in \mathbb{C}$ and a nonzero tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ are called eigenvalue and eigentensor of the tensor $\mathcal{A} \in \mathbb{T}$ if they satisfy the multilinear equation $\mathcal{A} \bar{\times}_N \mathcal{X} = \lambda \mathcal{X}$. We define the spectral radius $\rho(\mathcal{A})$ of $\mathcal{A}$ to be the largest modulus of the eigenvalues of $\mathcal{A}$. We define a stochastic process to be weakly stationary if the first and second moment of its finite dimensional distributions are finite and constant in t. Finally, note that it is always possible to rewrite an ART(p) process as a ART(1) process on an augmented state space, by stacking the endogenous tensors along the first mode. Thus, without loss of generality, we focus on the case $p = 1$. We use the definition of inverse tensor, spectral radius, and the convergence of power series of tensors to prove the following results (see Section S.4 in the supplement for the proofs).

Lemma 2.1. *Every $(I_1 \times I_2 \times \ldots \times I_N \times I_1 \times I_2 \times \ldots \times I_N)$-dimensional ART(p) process $\mathcal{Y}_t = \sum_{k=1}^{p} \mathcal{A}_k \bar{\times}_N \mathcal{Y}_{t-j} + \mathcal{E}_t$, $t \in \mathbb{Z}$, can be rewritten as a $(pI_1 \times I_2 \times \ldots \times I_N \times pI_1 \times I_2 \times \ldots \times I_N)$-dimensional ART(1) process $\underline{\mathcal{Y}}_t = \underline{\mathcal{A}} \bar{\times}_N \underline{\mathcal{Y}}_{t-1} + \underline{\mathcal{E}}_t$, $t \in \mathbb{Z}$.*

Proposition 2.1 (Stationarity)**.** *If $\rho(\widetilde{\mathcal{A}}_1) < 1$ and the process $\mathcal{X}_t$, $t \in \mathbb{Z}$, is weakly stationary, then the ART process in Eq. (3), with $p = 1$, is weakly stationary and admits the representation*

$$\mathcal{Y}_t = (\mathcal{I} - \widetilde{\mathcal{A}}_1)^{-1} \bar{\times}_N \widetilde{\mathcal{A}}_0 + \sum_{k=0}^{\infty} \widetilde{\mathcal{A}}_1^k \bar{\times}_N \widetilde{\mathcal{B}} \bar{\times}_M \mathcal{X}_{t-k} + \sum_{k=0}^{\infty} \widetilde{\mathcal{A}}_1^k \bar{\times}_N \mathcal{E}_{t-k}, \qquad t \in \mathbb{Z}.$$

Proposition 2.2. *The VAR(p) in Eq. (16) is weakly stationary if and only if the ART(p) in Eq. (3) is weakly stationary.*

2.3 Parametrization

The unrestricted model in Eq. (5) cannot be estimated, as the number of parameters greatly outmatches the available data. We address this issue by assuming a PARAFAC(R) decomposition for the tensor coefficients, which makes the estimation feasible by reducing the dimension of the parameter space. The models in Eqq. (5)-(3) are equivalent but the assuming a PARAFAC decomposition for the coefficient tensors leads to different degrees of parsimony, as shown in the following remark.

Remark 2.2 (Parametrization via contracted product)**.** *The two models* (5) *and* (3) *combined with the PARAFAC decomposition for the tensor coefficients allow for different degree of parsimony. To show this, without loss of generality, focus on the coefficient tensor* $\widetilde{\mathcal{A}}_1$ *(similar argument holds for* $\widetilde{\mathcal{A}}_j$, $j = 2, \ldots, p$ *and* $\widetilde{\mathcal{B}}$*). By assuming a PARAFAC(R) decomposition for* $\widetilde{\mathcal{A}}_1$ *in* (3) *and for* $\mathcal{A}_1$ *in* (5)*, we get, respectively*

$$\widetilde{\mathcal{A}}_1 = \sum_{r=1}^{R} \widetilde{\boldsymbol{\alpha}}_1^{(r)} \circ \ldots \circ \widetilde{\boldsymbol{\alpha}}_N^{(r)} \circ \widetilde{\boldsymbol{\alpha}}_{N+1}^{(r)} \circ \ldots \circ \widetilde{\boldsymbol{\alpha}}_{2N}^{(r)}, \quad \mathcal{A}_1 = \sum_{r=1}^{R} \boldsymbol{\alpha}_1^{(r)} \circ \ldots \circ \boldsymbol{\alpha}_N^{(r)} \circ \boldsymbol{\alpha}_{N+1}^{(r)},$$

The length of the vectors $\boldsymbol{\alpha}_j^{(r)}$ *and* $\widetilde{\boldsymbol{\alpha}}_j^{(r)}$ *coincide for each* $j = 1, \ldots, N$*. However,* $\boldsymbol{\alpha}_{N+1}^{(r)}$ *has length* I^* *while* $\widetilde{\boldsymbol{\alpha}}_{N+1}^{(r)}, \ldots, \widetilde{\boldsymbol{\alpha}}_{2N}^{(r)}$ *have length* $I_1, \ldots, I_N$*, respectively. Therefore, the number of free parameters in the coefficient tensor* $\mathcal{A}_1$ *is* $R(I_1 + \ldots + I_N + \prod_{j=1}^{N} I_j)$*, while it is* $2R(I_1 + \ldots + I_N)$ *for* $\widetilde{\mathcal{A}}_1$*. This highlights the greater parsimony granted by the use of the PARAFAC(R) decomposition in model* (3) *as compared to model* (5)*.*

Remark 2.3 (Vectorization)**.** *There is a relation between the* $(I_1 \times \ldots \times I_N)$*-dimensional ART(p) and a* $(I_1 \cdot \ldots \cdot I_N)$*-dimensional VAR(p) model. The vector form of* (5) *is*

$$\begin{aligned} \operatorname{vec}\left(\mathcal{Y}_t\right) &= \operatorname{vec}\left(\mathcal{A}_0\right) + \sum_{j=1}^{p} \operatorname{mat}_{N+1}(\mathcal{A}_j) \operatorname{vec}\left(\mathcal{Y}_{t-j}\right) + \operatorname{mat}_{N+1}(\mathcal{B}) \operatorname{vec}\left(\mathcal{X}_t\right) + \operatorname{vec}\left(\mathcal{E}_t\right) \\ \mathbf{y}_t &= \boldsymbol{\alpha}_0 + \sum_{j=1}^{p} \mathbf{A}'_{(N+1),j} \mathbf{y}_{t-j} + \mathbf{B}'_{(N+1)} \mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}_{I^*}(\mathbf{0}, \Sigma_N \otimes \ldots \otimes \Sigma_1), \end{aligned} \tag{6}$$

$t \in \mathbb{Z}$*, where the constraint on the covariance matrix stems from the one-to-one relation between the tensor normal distribution for* $\mathcal{X}$ *and the distribution of its vectorization (Ohlson et al., 2013) given by* $\mathcal{X} \sim \mathcal{N}_{I_1,\ldots,I_N}(\mathcal{M}, \Sigma_1, \ldots, \Sigma_N)$ *if and only if* $\operatorname{vec}\left(\mathcal{X}\right) \sim \mathcal{N}_{I^*}(\operatorname{vec}\left(\mathcal{M}\right), \Sigma_N \otimes \ldots \otimes \Sigma_1)$*. The restriction on the covariance structure for the vectorized tensor provides a parsimonious parametrization of the multivariate normal distribution, while allowing both within and between mode dependence. Alternative parametrizations for the covariance lead to generalizations of standard models. For example, assuming an additive covariance structure results in the tensor ANOVA. This is an active field for further research.*

Example 2.1. *For the sake of exposition, consider the model in Eq.* (5)*, where* $p = 1$*, the response is a 3-order tensor* $\mathcal{Y}_t \in \mathbb{R}^{d \times d \times d}$ *and the covariates include only a constant*

(a) vectorized (b) contracted product $\bar{\times}_N$ as (5) (c) contracted product $\bar{\times}_N$ as (3)

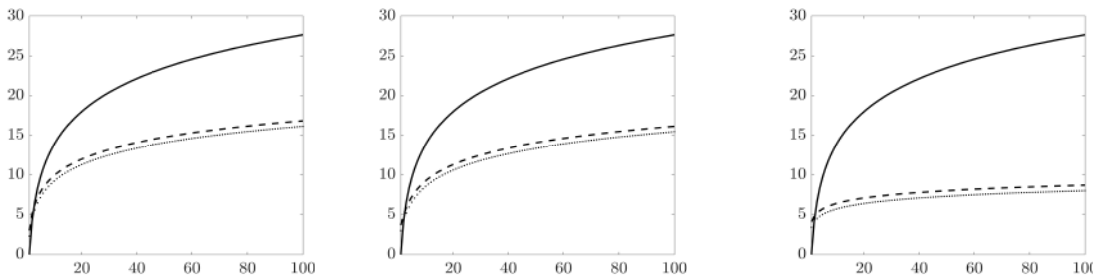

Figure 1: Number of parameters in $\mathcal{A}_0$, in log-scale (*vertical axis*) as function of the size d of the $(d \times d \times d)$-dimensional tensor $\mathcal{Y}_t$ (*horizontal axis*) in a ART(1) model. In all plots: unconstrained model (*solid* line), PARAFAC(R) parametrization with $R = 10$ (*dashed* line) and $R = 5$ (*dotted* line). Parametrizations: vectorized model (panel a), mode-n product of (5) (panel b) and contracted product of (3) (panel c).

coefficient tensor $\mathcal{A}_0$. Define by $k_\mathcal{E}$ the number of parameters of the noise distribution. The total number of parameters to estimate in the unrestricted case is $(d^{2N}) + k_\mathcal{E} = O(d^{2N})$, with $N = 3$ in this example. Instead, in a ART model defined via the mode-n product in Eq. (5), assuming a PARAFAC(R) decomposition on $\mathcal{A}_0$ the total number of parameters is $\sum_{r=1}^{R}(d^N + d^N) + k_\mathcal{E} = O(d^N)$. Finally, in the ART model defined by the contracted product in Eq. (3) with a PARAFAC(R) decomposition on $\widetilde{\mathcal{A}}_0$ the number of parameters is $\sum_{r=1}^{R} Nd + k_\mathcal{E} = O(d)$. A comparison of the different parsimony granted by the PARAFAC decomposition in all models is illustrated in Fig. 1.

The structure of the PARAFAC decomposition poses an identification problem for the marginals $\boldsymbol{\beta}_j^{(r)}$, which may arise from three sources:

(i) *scale* identification, since $\lambda_{jr}\boldsymbol{\beta}_j^{(r)} \circ \lambda_{kr}\boldsymbol{\beta}_k^{(r)} = \boldsymbol{\beta}_j^{(r)} \circ \boldsymbol{\beta}_k^{(r)}$ for any collection $\{\lambda_{jr}\}_{j,r}$ such that $\prod_{j=1}^{J} \lambda_{jr} = 1$;

(ii) *permutation* identification, since $\boldsymbol{\beta}_j^{(\pi(r))} \circ \boldsymbol{\beta}_k^{(\pi(r))} = \boldsymbol{\beta}_j^{(r)} \circ \boldsymbol{\beta}_k^{(r)}$ for any permutation π of the indices $\{1, \dots, R\}$;

(iii) *orthogonal transformation* identification, since $\boldsymbol{\beta}_j^{(r)}Q \circ \boldsymbol{\beta}_k^{(r)}Q = \boldsymbol{\beta}_j^{(r)}Q(\boldsymbol{\beta}_k^{(r)}Q)' = \boldsymbol{\beta}_j^{(r)} \circ \boldsymbol{\beta}_k^{(r)}$ for any orthonormal matrix Q.

Note that in our framework these issues do not hamper the inference, since our object of interest is the coefficient tensor $\mathcal{B}$, which is exactly identified. The marginals $\boldsymbol{\beta}_j^{(r)}$ have no interpretation, as the PARAFAC decomposition is assumed on the coefficient tensor for the sake of providing a parsimonious parametrization.

2.4 Important Special Cases

The model in Eq. (5) is a generalization of several well-known linear econometric models, such as univariate regression, VARX, SUR, panel VAR, VECM and matrix autoregressive models (MAR). See Sections S.3-S.4 of the supplement for further details. Dropping the covariates $\mathcal{X}_t$ from Eq. (5), we obtain an autoregressive tensor model of order p (or ART(p))

$$\mathcal{Y}_t = \mathcal{A}_0 + \sum_{j=1}^{p} \mathcal{A}_j \times_{N+1} \operatorname{vec}\left(\mathcal{Y}_{t-j}\right) + \mathcal{E}_t, \qquad \mathcal{E}_t \overset{iid}{\sim} \mathcal{N}_{I_1,\ldots,I_N}(\mathbf{0}, \Sigma_1, \ldots, \Sigma_N), \qquad t \in \mathbb{Z}. \quad (7)$$

2.5 Impulse Response Analysis

In this section we derive two impulse response functions (IRF) for ART models, the block Cholesky IRF and the block generalised IRF, exploiting the relationship between ART and VAR models. Without loss of generality, we focus on the ART(p) model in Eq. (7), with $p = 1$ and $\mathcal{A}_0 = \mathbf{0}$, and introduce the following notation. Let $\mathbf{y}_t = \operatorname{vec}\left(\mathcal{Y}_t\right)$ and $\boldsymbol{\epsilon}_t = \operatorname{vec}\left(\mathcal{E}_t\right) \sim \mathcal{N}_{I^*}(\mathbf{0}, \Sigma)$ be the $(I^* \times 1)$ tensor response and noise term in vector form, respectively, where $\Sigma = \Sigma_N \otimes \ldots \otimes \Sigma_1$ is the $(I^* \times I^*)$ covariance of the model in vector form and $I^* = \prod_{k=1}^{N} I_k$. Partition Σ in blocks as

$$\Sigma = \left(\begin{array}{c|c} A & B \\ \hline B' & C \end{array}\right), \quad (8)$$

where A is $n \times n$, B is $n \times (I^* - n)$, and C is $(I^* - n) \times (I^* - n)$, with $1 \leq n \leq I^*$. Then, denoting by $S = C - B'A^{-1}B$ the Schur complement of A, the LDU decomposition of Σ is

$$\Sigma = \left(\begin{array}{c|c} \mathbf{I}_n & \mathrm{O}_{n,I^*-n} \\ \hline B'A^{-1} & \mathbf{I}_{I^*-n} \end{array}\right) \left(\begin{array}{c|c} A & \mathrm{O}_{n,I^*-n} \\ \hline \mathrm{O}'_{n,I^*-n} & S \end{array}\right) \left(\begin{array}{c|c} \mathbf{I}_n & A^{-1}B \\ \hline \mathrm{O}'_{n,I^*-n} & \mathbf{I}_{I^*-n} \end{array}\right) = LDL'.$$

where $\mathbf{I}_j$ is the identity matrix of size j. Hence Σ can be block-diagonalised

$$D = L^{-1}\Sigma(L')^{-1} = \left(\begin{array}{c|c} A & \mathrm{O}_{n,I^*-n} \\ \hline \mathrm{O}'_{n,I^*-n} & S \end{array}\right). \quad (9)$$

From the Cholesky decomposition of D one obtains a block Cholesky decomposition

$$\Sigma = \left(\begin{array}{c|c} L_A & \mathrm{O}_{n,I^*-n} \\ \hline B'(L_A^{-1})' & L_S \end{array}\right) \left(\begin{array}{c|c} L'_A & L_A^{-1}B \\ \hline \mathrm{O}'_{n,I^*-n} & L'_S \end{array}\right) = PP',$$

where L_A, L_S are the Cholesky factors of A and S, respectively. Assume the vectorised ART process admits an infinite MA representation, with $\Psi_0 = \mathbf{I}_{I^*}$ and $\Psi_i = \mathrm{mat}_{(4)}(\mathcal{B})'\Psi_{i-1}$, then using the previous results we get:

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \Psi_i \boldsymbol{\epsilon}_{t-i} = \sum_{i=0}^{\infty} (\Psi_i L)(L^{-1}\boldsymbol{\epsilon}_{t-i}) = \sum_{i=0}^{\infty} (\Psi_i L)\boldsymbol{\eta}_{t-i} \quad \boldsymbol{\eta}_t \sim \mathcal{N}_{I^*}(\mathbf{0}, D), \tag{10}$$

where $\boldsymbol{\eta}_t = L^{-1}\boldsymbol{\epsilon}_t$ are the block-orthogonalised shocks and D is the block-diagonal matrix in Eq. (9). Denote with E_n the $I^* \times n$ matrix that selects n columns from a pre-multiplied matrix, i.e. DE_n is a matrix containing n columns of D. Denote with $\boldsymbol{\delta}^*$ a n-dimensional vector of shocks. Using the property of the multivariate Normal distribution, and recalling that the top-left block of size n of D is A, we extend the generalised IRF of Koop et al. (1996) and Pesaran and Shin (1998) by defining the block generalised IRF

$$\begin{aligned}\boldsymbol{\psi}^G(h;n) &= \mathbb{E}\big(\operatorname{vec}\big(\mathcal{Y}_{t+h}\big)|\operatorname{vec}\big(\mathcal{E}_t\big)' = (\boldsymbol{\delta}^{*\prime}, \mathbf{0}'_{I^*-n}), \mathcal{F}_{t-1}\big) - \mathbb{E}\big(\operatorname{vec}\big(\mathcal{Y}_{t+h}\big)|\mathcal{F}_{t-1}\big) \\ &= (\Psi_h L) D E_n A^{-1} \boldsymbol{\delta}^*, \qquad h = 1, 2, \ldots \end{aligned} \tag{11}$$

where $\mathcal{F}_u$, $u \leq t$ is the natural filtration associated to the stochastic process $\mathcal{Y}_t$, $t \in \mathbb{Z}$. Starting from Eq. (10) we derive the block Cholesky IRF (OIRF) as

$$\begin{aligned}\boldsymbol{\psi}^O(h;n) &= \mathbb{E}\big(\operatorname{vec}\big(\mathcal{Y}_{t+h}\big)|\operatorname{vec}\big(\mathcal{E}_t\big)' = (\boldsymbol{\delta}^{*\prime}, \mathbf{0}'_{I^*-n}), \mathcal{F}_{t-1}\big) \\ &\quad - \mathbb{E}\big(\operatorname{vec}\big(\mathcal{Y}_{t+h}\big)|\operatorname{vec}\big(\mathcal{E}_t\big)' = \mathbf{0}'_{I^*}, \mathcal{F}_{t-1}\big) \\ &= (\Psi_h L) P E_n \boldsymbol{\delta}^*, \qquad h = 1, 2, \ldots . \end{aligned} \tag{12}$$

Define with $\mathbf{e}_j$ the j-th column of the I^*-dimensional identity matrix. The impact of a shock δ^* to the j-th variable on all I^* variables is given below in Eq. (13), whereas the impact of a shock to the j-th variable on the i-th variable is given in Eq. (14).

$$\boldsymbol{\psi}_j^G(h;n) = \Psi_h L D \mathbf{e}_j D_{jj}^{-1} \delta^*, \qquad \boldsymbol{\psi}_j^O(h;n) = \Psi_h L P \mathbf{e}_j \delta^* \tag{13}$$

$$\psi_{ij}^G(h;n) = \mathbf{e}_i' \Psi_h L D \mathbf{e}_j D_{jj}^{-1} \delta^*, \qquad \psi_{ij}^O(h;n) = \mathbf{e}_i' \Psi_h L P \mathbf{e}_j \delta^*. \tag{14}$$

Finally, denoting $\boldsymbol{\delta}_j = \mathbf{e}_j \delta^*$, we have the compact notation

$$\begin{aligned}\boldsymbol{\psi}_j^G(h;n) &= \Psi_h L D D_{jj}^{-1} \boldsymbol{\delta}_j, \qquad & \boldsymbol{\psi}_j^O(h;n) &= \Psi_h L P \boldsymbol{\delta}_j \\ \psi_{ij}^G(h;n) &= \mathbf{e}_i' \Psi_h L D D_{jj}^{-1} \boldsymbol{\delta}_j, \qquad & \psi_{ij}^O(h;n) &= \mathbf{e}_i' \Psi_h L P \boldsymbol{\delta}_j. \end{aligned}$$

3 Bayesian Inference

In this section, without loss of generality, we present the inference procedure for a special case of the model in Eq. (5), given by

$$\mathcal{Y}_t = \mathcal{B} \times_4 \operatorname{vec}\left(\mathcal{Y}_{t-1}\right) + \mathcal{E}_t, \qquad \mathcal{E}_t \overset{iid}{\sim} \mathcal{N}_{I_1,I_2,I_3}(\mathbf{0}, \Sigma_1, \Sigma_2, \Sigma_3). \tag{15}$$

Here $\mathcal{Y}_t$ is a 3-order tensor response of size $I_1 \times I_2 \times I_3$, $\mathcal{X}_t = \mathcal{Y}_{t-1}$ and $\mathcal{B}$ is thus a 4-order coefficient tensor of size $I_1 \times I_2 \times I_3 \times I_4$, with $I_4 = I_1 I_2 I_3$. This is a 3-order *tensor autoregressive model* of lag-order 1, or ART(1), coinciding with Eq. (7) for $p = 1$ and $\mathcal{A}_0 = \mathbf{0}$. The noise term $\mathcal{E}_t$ has as tensor normal distribution, with zero mean and covariance matrices $\Sigma_1, \Sigma_2, \Sigma_3$ of sizes $I_1 \times I_1$, $I_2 \times I_2$ and $I_3 \times I_3$, respectively, accounting for the covariance along each of the three dimensions of $\mathcal{Y}_t$. The specification of a tensor model with a tensor normal noise instead of a vector model (like a Gaussian VAR) has the advantage of being more parsimonious. By vectorising (15), we get the equivalent VAR

$$\operatorname{vec}\left(\mathcal{Y}_t\right) = \mathbf{B}'_{(4)} \operatorname{vec}\left(\mathcal{Y}_{t-1}\right) + \operatorname{vec}\left(\mathcal{E}_t\right), \quad \operatorname{vec}\left(\mathcal{E}_t\right) \overset{iid}{\sim} \mathcal{N}_{I^*}(\mathbf{0}, \Sigma_3 \otimes \Sigma_2 \otimes \Sigma_1), \tag{16}$$

whose covariance has a Kronecker structure, which contains $(I_1(I_1 + 1) + I_2(I_2 + 1) + I_3(I_3 + 1))/2$ parameters (as opposed to $(I^*(I^* + 1))/2$ of an unrestricted VAR) and allows for heteroskedasticity.

The choice the Bayesian approach for inference is motivated by the fact that the large number of parameters may lead to an overfitting problem, especially when the samples size is rather small. This issue can be addressed by the indirect inclusion of parameter restrictions through a suitable specification of the corresponding prior distributions. In the unrestricted model (15) it would be necessary to define a prior distribution on the 4-order tensor $\mathcal{B}$. The literature on tensor-valued distributions is limited to the elliptical family (e.g. Ohlson et al., 2013), which includes the tensor normal and tensor t. Both distributions do not easily allow for the specification of restrictions on a subset of the entries of the tensor, hampering the use of standard regularization prior distributions (such as shrinkage priors).

The PARAFAC(R) decomposition of the coefficient tensor provides a way to circumvent this issue. This decomposition allows to represent a tensor through a collection of vectors (the marginals), for which many flexible shrinkage prior distributions are available. Indirectly, this introduces *a priori* shrinkage to zero of the coefficient tensor.

3.1 Prior Specification

The choice of the prior distribution on the PARAFAC marginals is crucial for shrinking towards zero some elements of the coefficient tensor and for increasing the efficiency of the inference. Global-local prior distributions are based on scale mixtures of normal distributions, where the different components of the covariance matrix govern the amount of prior shrinkage. Compared to spike-and-slab distributions (e.g. George and McCulloch, 1997; Ishwaran and Rao, 2005; Mitchell and Beauchamp, 1988) which become infeasible as the parameter space grows, global-local priors have better scalability properties in high-dimensional settings. They do not provide automatic variable selection, which can nonetheless be obtained by post-estimation thresholding (Park and Casella, 2008).

Motivated by these arguments, we define a global-local shrinkage prior for the marginals $\boldsymbol{\beta}_j^{(r)}$ of the coefficient tensor $\mathcal{B}$ following the hierarchical prior specification of Guhaniyogi et al. (2017). For each $\boldsymbol{\beta}_j^{(r)}$, we define a prior distributions as a scale mixture of normals centred in zero, with three components for the covariance. The global parameter τ governs the overall variance, the middle parameter ϕ_r defines the common shrinkage for the marginals in r-th component of the PARAFAC, and the local parameter $W_{j,r} = \mathrm{diag}(\mathbf{w}_{j,r})$ drives the shrinkage of each entry of each marginal. Summarizing, for $p = 1, \dots, I_j$, $j = 1, \dots, J$ ($J = 4$ in Eq. (15)) and $r = 1, \dots, R$, the hierarchical prior structure (we use the shape-rate formulation for the gamma distribution) for each vector of the PARAFAC(R) decomposition in Eq. (1) is

$$\begin{aligned}
\pi(\boldsymbol{\phi}) \sim \mathcal{D}ir(\alpha \mathbf{1}_R) \quad \pi(\tau) \sim \mathcal{G}a(a_\tau, b_\tau) \quad \pi(\lambda_{j,r}) \sim \mathcal{G}a(a_\lambda, b_\lambda) \\
\pi(w_{j,r,p} | \lambda_{j,r}) \sim \mathcal{E}xp(\lambda_{j,r}^2/2) \\
\pi\big(\boldsymbol{\beta}_j^{(r)} \big| W_{j,r}, \boldsymbol{\phi}, \tau\big) \sim \mathcal{N}_{I_j}(\mathbf{0}, \tau \phi_r W_{j,r}),
\end{aligned} \tag{17}$$

where $\mathbf{1}_R$ is the vector of ones of length R and we assume $a_\tau = \alpha R$ and $b_\tau = \alpha R^{1/J}$. The conditional prior distribution of a generic entry $b_{i_1,\dots,i_J}$ of $\mathcal{B}$ is the law of a sum of product Normals (a product Normal is the distribution of the product of n independent centred Normal random variables): it is symmetric around zero, with fatter tails than both a standard Gaussian or a standard Laplace distribution (see Section S.5 of the supplement for further details). The peak at zero of the product Normal prior promotes shrinking effects. The following result characterises the conditional prior distribution of an entry of the coefficient tensor $\mathcal{B}$ induced by the hierarchical prior in Eq. (17). See Section S.5 for the proof.

Lemma 3.1. *Let* $b_{ijkp} = \sum_{r=1}^{R} \beta_r$*, where* $\beta_r = \beta_{1,i}^{(r)}\beta_{2,j}^{(r)}\beta_{3,k}^{(r)}\beta_{4,p}^{(r)}$*, and let* $m_1 = i$*,* $m_2 = j$*,* $m_3 = k$ *and* $m_4 = p$*. Under the prior specification in (17), the generic entry* b_{ijkp} *of the coefficient tensor* $\mathcal{B}$ *has the conditional prior distribution*

$$\pi(b_{ijkp}|\tau, \boldsymbol{\phi}, \mathbf{W}) = p\left(\sum_{r=1}^{R} \beta_r \Big| -\right) = p(\beta_1|-) * \ldots * p(\beta_R|-),$$

where $*$ *denotes the convolution and*

$$p(\beta_r|-) = K_r \cdot G_{4,0}^{4,0}\left(\beta_r^2 \prod_{h=1}^{4}(2\tau\phi_r w_{h,r,m_h})^{-1}\Big|\mathbf{0}\right),$$

with $G_{p,q}^{m,n}(x|\ {\mathbf{a} \atop \mathbf{b}})$ *a Meijer G-function and*

$$G_{4,0}^{4,0}\left(\beta_r^2 \prod_{h=1}^{4}(2\tau\phi_r w_{h,r,m_h})^{-1}\Big|\mathbf{0}\right) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\left(\beta_r^2 \prod_{h=1}^{4}(2\tau\phi_r w_{h,r,m_h})^{-1}\right)^{-s} \mathrm{d}s$$

$$K_r = (2\pi)^{-4/2} \prod_{h=1}^{4}(2\tau\phi_r w_{h,r,m_h})^{-1}.$$

The use of Meijer G-functions and Fox H-functions is not new in econometrics. They arise as limiting distributions for the cointegrating vector in VECM models (e.g., Abadir and Paruolo, 1997) and have been used for defining prior distributions in Bayesian analysis of non-conjugate Gaussian models (Andrade and Rathie, 2015, 2017).

From Eq. (4), we have that the covariance matrices Σ_j, $j = 1, \ldots, J$, enter the likelihood in a multiplicative way, therefore separate identification of their scales requires further restrictions. Wang and West (2009) and Dobra (2015) adopt independent hyper-inverse Wishart prior distributions (Dawid and Lauritzen, 1993) for each Σ_j, then impose the identification restriction $\Sigma_{j,11} = 1$ for $j = 2, \ldots, J-1$. The hard constraint $\Sigma_j = \mathbf{I}_{I_j}$, for all but one n, implicitly imposes that the dependence structure within different modes is the same, but there is no dependence between modes. We follow Hoff (2011), who suggests to introduce dependence between the Inverse Wishart prior distribution of each Σ_j via a hyper-parameter γ affecting their prior scale. To account for marginal dependence, we add a level of hierarchy, thus obtaining

$$\pi(\gamma) \sim \mathcal{G}a(a_\gamma, b_\gamma) \qquad \pi(\Sigma_j|\gamma) \sim \mathcal{IW}_{I_j}(\nu_j, \gamma\Psi_j). \tag{18}$$

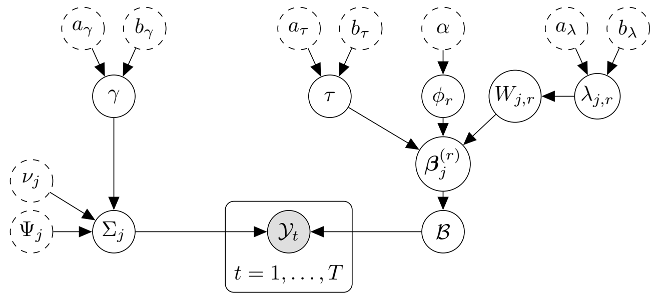

Figure 2: Directed acyclic graph of the model in Eq. (15) and prior structure in Eqq. (17)-(18). Gray circles denote observable variables, white solid circles indicate parameters, white dashed circles indicate fixed hyperparameters. Directed edges represent the conditional independence relationships.

Define $\Lambda = \{\lambda_{j,r} : j = 1, \ldots, J,\ r = 1, \ldots, R\}$ and $\mathbf{W} = \{W_{j,r} : j = 1, \ldots, J,\ r = 1, \ldots, R\}$, and let $\boldsymbol{\theta}$ denote the collection of all parameters. The directed acyclic graph (DAG) of the prior structure is given in Fig. 2.

Note that our prior specification is flexible enough to include Minnesota-type restrictions or hierarchical structures as in Canova and Ciccarelli (2004).

3.2 Posterior Computation

Define $\mathbf{Y} = \{\mathcal{Y}_t\}_{t=1}^T$, $I_0 = \sum_{j=1}^J I_j$, $\boldsymbol{\beta}_{-j}^{(r)} = \{\boldsymbol{\beta}_i^{(r)} : i \neq j\}$ and $\mathcal{B}_{-r} = \{B_i : i \neq r\}$, with $B_r = \boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_4^{(r)}$. The likelihood function of model (15) is

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{t=1}^T (2\pi)^{-\frac{I_4}{2}} \prod_{j=1}^3 \left|\Sigma_j\right|^{-\frac{I_{-j}}{2}} \tag{19}$$
$$\cdot \exp\Big(-\frac{1}{2}\Sigma_2^{-1}(\mathcal{Y}_t - \mathcal{B} \times_4 \mathbf{y}_{t-1}) \times_{1\ldots3}^{1\ldots3} \left(\circ_{j=1}^3 \Sigma_j^{-1}\right) \times_{1\ldots3}^{1\ldots3} (\mathcal{Y}_t - \mathcal{B} \times_4 \mathbf{y}_{t-1})\Big),$$

where $\mathbf{y}_{t-1} = \text{vec}\left(\mathcal{Y}_{t-1}\right)$ and $\boldsymbol{\theta}$ denotes the collection of all parameters. Since the posterior distribution is not tractable, we adopt an MCMC procedure based on Gibbs sampling. The details of the derivation of the full conditional posterior distributions are given in Section S.6 of the supplement. We articulate the sampler in three main blocks:

(I) Sample the global and middle variance hyper-parameters of the marginals, from

$$p(\psi_r|\mathcal{B}, \mathbf{W}, \alpha) \propto \text{GiG}\big(\alpha - I_0/2,\ 2b_\tau,\ 2C_r\big) \tag{20}$$

$$p(\tau|\mathcal{B}, \mathbf{W}, \boldsymbol{\phi}) \propto \text{GiG}\Big(a_\tau - RI_0/2,\ 2b_\tau,\ 2\sum_{r=1}^R C_r/\phi_r\Big), \tag{21}$$

where $C_r = \sum_{j=1}^{J} \boldsymbol{\beta}_j^{(r)'} W_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)}$, then set $\phi_r = \psi_r / \sum_{l=1}^{R} \psi_l$. To improve the mixing, we sample τ with a Hamiltonian Monte Carlo (HMC) step (Neal, 2011).

(II) Sample the local variance hyper-parameters of the marginals and the marginals themselves, from

$$p\big(\lambda_{j,r} | \boldsymbol{\beta}_j^{(r)}, \phi_r, \tau\big) \propto \mathcal{G}a\big(a_\lambda + I_j,\, b_\lambda + \big\|\boldsymbol{\beta}_j^{(r)}\big\|_1 (\tau\phi_r)^{-1/2}\big) \tag{22}$$

$$p\big(w_{j,r,p} | \lambda_{j,r}, \phi_r, \tau, \boldsymbol{\beta}_j^{(r)}\big) \propto \mathrm{GiG}\big(1/2,\, \lambda_{j,r}^2,\, (\beta_{j,p}^{(r)})^2/(\tau\phi_r)\big) \tag{23}$$

$$p\big(\boldsymbol{\beta}_j^{(r)} | \boldsymbol{\beta}_{-j}^{(r)}, \mathcal{B}_{-r}, W_{j,r}, \phi_r, \tau, \mathbf{Y}, \Sigma_1, \ldots, \Sigma_3\big) \propto \mathcal{N}_{I_j}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_j}, \bar{\Sigma}_{\boldsymbol{\beta}_j}). \tag{24}$$

(III) Sample the covariance matrices and the latent scale, from

$$p(\Sigma_j | \mathcal{B}, \mathbf{Y}, \Sigma_{-j}, \gamma) \propto \mathcal{IW}_{I_j}(\nu_j + I_j,\, \gamma\Psi_j + S_j) \tag{25}$$

$$p(\gamma | \Sigma_1, \ldots, \Sigma_3) \propto \mathcal{G}a\Big(a_\gamma + \sum_{j=1}^{3} \nu_j I_j,\, b_\gamma + \sum_{j=1}^{3} \mathrm{tr}(\Psi_j \Sigma_j^{-1})\Big). \tag{26}$$

4 Application to Multilayer Dynamic Networks

We apply the proposed methodology to study jointly the dynamics of international trade and credit networks. The international trade network has been previously investigated by several authors (e.g., Eaton and Kortum, 2002; Fieler, 2011), but to the best of our knowledge, this is the first attempt to model the dynamics of two networks jointly. Moreover, the impulse response analysis in this setting can be used for predicting possible trade creation and diversion effects (e.g., Bikker, 2010).

The bilateral trade data come from the COMTRADE database, whereas the data on bilateral outstanding credit come from the Bank of International Settlements database. Our sample of yearly observations for 10 countries runs from 2003 to 2016. At each time t, the 3-order tensor $\mathcal{Y}_t$ has size $(10, 10, 2)$ and represents a 2-layer node-aligned network (or multiplex) with 10 vertices (countries), where each edge is given by a bilateral trade flow or financial exposure. See Section S.9 in the supplement for data description.

We estimate the tensor autoregressive model in Eq. (15), using the prior structure described in Section 3, and run the Gibbs sampler for $N = 100,000$ iterations after $30,000$ burn-in iterations. We retain every second draw for posterior inference.

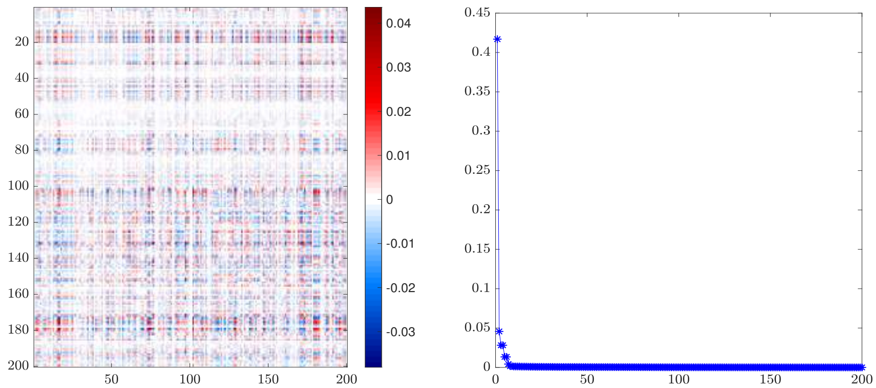

Figure 3: *Left:* mode-4 matricization of estimated coefficient tensor $\hat{B}_{(4)}$. *Right:* log-spectrum of $\hat{B}_{(4)}$, decreasing order.

The mode-4 matricization of the estimated coefficient tensor, $\hat{B}_{(4)}$, is shown in the left panel of Fig. 3. The (i, j)-th entry of the matrix $\hat{B}_{(4)}$ reports the impact of the edge j on edge i in vectorised form (e.g., $j = 21$ and $i = 4$ corresponds to the coefficient of entry $\mathcal{Y}_{1,3,1,t-1}$ on $\mathcal{Y}_{4,1,1,t}$). The first 100 rows/columns correspond to the edges in the first layer. Hence, two rows of the matricized coefficient tensor are similar when two edges are affected by all the edges of the (lagged) network in a similar way, whereas two similar columns identify the situation where two edges impact the (next period) network in a similar way. The overall distribution of the estimated entries of $\hat{B}_{(4)}$ is symmetric around zero and leptokurtic, as a consequence of the shrinkage to zero of the estimated coefficients. The right panel of Fig. 3 shows the log-spectrum of $\hat{B}_{(4)}$. As all eigenvalues of $\hat{B}_{(4)}$ have modulus smaller than one, we conclude that the estimated ART(1) model is weakly stationary. In fact, it can be shown that the stationarity of the mode-4 matricised coefficient tensor implies stationarity of the ART(1) process. Additional estimation results are provided in Section S.10 of the supplement.

After estimating the ART(1) model (15), we may investigate shock propagation across the network computing generalised and orthogonalised impulse response functions presented in equations (11) and (12), respectively. Impulse responses allow us to analyze the propagation of shocks both across the network, within and across layers, and over time. For illustration, we study the responses to a shock in all edges of a country, by applying block Cholesky factorisation to Σ, in such a way that the shocked country contemporaneously affects all others and not vice-versa (we do not report generalised IRFs, which are very similar). Thus, the matrices A and C in Eq. (8) reflect contemporaneous correlations across transactions of the shock-originating country and with transactions of all other countries, respectively. For expositional convenience, we report only statistically significant responses.

In this analysis we consider a negative 1% shock to US trade imports (i.e., we allocate the shock across import originating countries to match import shares as in the last period of the sample). The results of the block Cholesky IRF at horizon 1 are given in Fig. 4. We report the impact on the whole network (panel (a)) and, for illustrative purposes, the impact on Germany's transactions (panel (b)).

Global effect on the network. The negative shock to US imports has an effect on both layers (trade and financial) of the network. There is evidence of heterogeneous responses across countries and country-specific transactions. On average, trade flows exhibit a slight expansion in response to the shock. Switzerland is the most positively affected, both in terms of exports and imports, and trade imports of the US show (on average) a reverted positive response one period after the shock. This reflects an oscillating impulse response. The overall average effect on the financial layer is negative, similar in magnitude to the effect on the trade layer. More specifically, we observe that Denmark's and Sweden's exports to Switzerland, Germany and France show a contraction, whereas the effect on US's, Japan's, and Ireland's exports to these countries is positive. We may interpret these effects as substitution effects: The decreasing share of Denmark's and Sweden's exports to Switzerland, Germany and France is offset by an increase in exports to the US, Japan and Ireland. In conclusion, model (15) permits to forecast trade creation and diversion effects (Bikker, 2010).

Local effect on Germany. In panel (b) of Fig. 4 we report the response of Germany's transactions to the negative shock in US imports. The effects on imports are mixed: while Germany's imports from most other EU countries increase, imports from Sweden and Denmark decrease. Likewise, Germany's exports show heterogeneous responses, whereby exports to Switzerland react strongest (positively). The shock in US imports does not have a significant impact on Germany's outstanding credit against most countries (except Switzerland and Japan). On the other hand, the reactions of Germany's outstanding debt reflect those on trade imports.

Local effect on other countries. We observe that the most affected trade transactions are those of Denmark, Japan, Ireland, Sweden and US (as exporters) vis-à-vis Switzerland and France (as importers). The financial layer mirrors these effects with opposite sign, while the magnitudes are comparable. Outstanding credit of Ireland and Japan to Switzerland, Germany and France decrease at horizon 1. By contrast, Denmark's outstanding credit to these countries increases. Note that outstanding debt of US vis-à-vis almost all countries decreases after the shock. Overall, responses to a shock on US imports at horizon 1 are heterogeneous in sign but rather low in magnitude, whereas at horizon 2 (plot not reported) the propagation of the shock has vanished. We interpret this as a sign of fast (and monotone) decay of the IRF.

(a) Network IRF at $h = 1$ (b) IRF for Germany's edges at $h = 1$

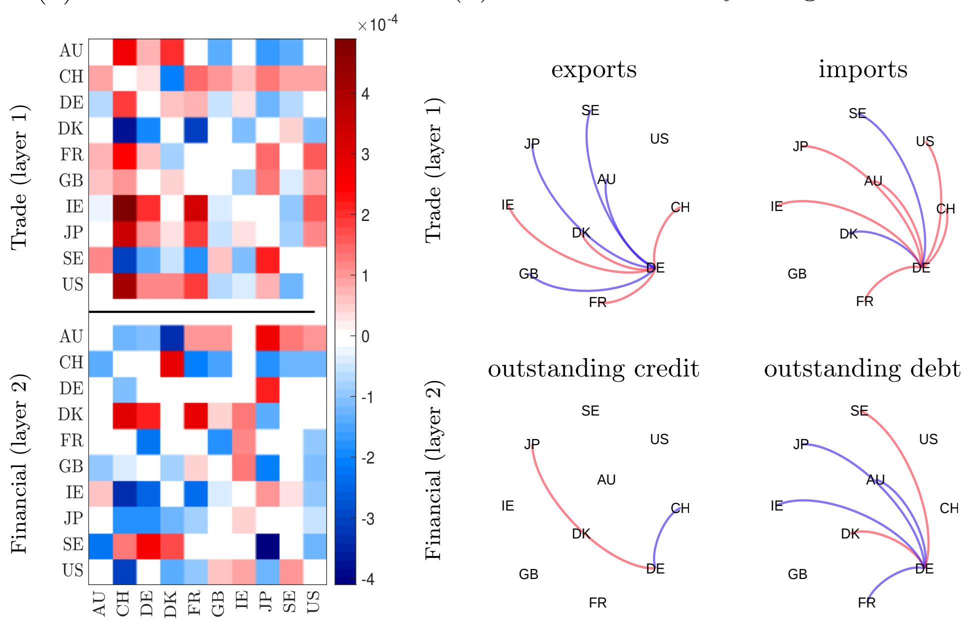

Figure 4: Shock to US trade imports by -1%. IRF at horizon $h = 1$ for all (*panel a*) and Germany (*panel b*) financial and trade transactions. In each plot negative coefficients are in blue and positive in red.

In addition, Section S.10 in the supplement shows additional impulse responses to a (i) negative 1% shock to Great Britain's (GB) outstanding debt and (ii) 1% negative shock to GB's outstanding debt coupled with a 1% positive shock to GB's outstanding credit.

5 Conclusions

We defined a new and general statistical framework for dynamic tensor regression. It encompasses the autoregressive tensor model, called ART, and many models frequently used in time series analysis as special cases, such as VAR, panel VAR, SUR, and MAR models. We exploited a low-rank decomposition of the coefficient tensor to reduce the parameter space dimension and specified a global-local shrinkage prior to address the overfitting. Taking advantage of the properties of the contracted product, we studied the main properties of the ART process and derived the impulse response function and the forecast error variance decomposition, which are essential tools for making predictions.

The proposed methodology has been applied to a time series of international trade and financial multilayer network. We are able to provide evidence of stationarity of the network process, heterogeneity in the shock propagation across countries and over time.

Acknowledgements

We are grateful to Federico Bassetti, Sylvia Frühwirth-Schnatter, Christian Gouriéroux, Søren Johansen, Siem Jan Koopman, Gary Koop, André Lucas, Alain Monfort, Peter Phillips, Raquel Prado, Christian P. Robert, Mark Steel, and Mike West for their comments and suggestions. Also, we thank the seminar participants at Queen Mary University, CREST, University of Warwick, University of Southampton, Vrije University of Amsterdam, London School of Economics, Maastricht University, and Polytechnic University of Milan. We thank the participants at "ES Annual Meeting" in Milan (2020), "ICEEE" in Lecce (2019), "CFENetwork" in Pisa (2018), "EC2" in Rome (2018), "RCEA Annual meeting" in Rimini (2018), "CFENetwork" in London (2017), "ICEEE" in Messina (2017), "Vienna Workshop on High-dimensional Time Series in Macroeconomics and Finance" in Wien (2017), "BISP10" in Milan (2017), "ESOBE" in Venice (2016), "CFENetwork" in Seville (2016), for their constructive comments.

Funding

This research used the SCSCF and HPC multiprocessor cluster systems and is part of the project Venice Center for Risk Analytics (VERA) at Ca' Foscari University of Venice. Monica Billio and Roberto Casarin acknowledge financial support from the Italian Ministry MIUR under the PRIN project Hi-Di NET - Econometric Analysis of High Dimensional Models with Network Structures in Macroeconomics and Finance (grant agreement no. 2017TA7TYC). Matteo Iacopini acknowledges financial support from the EU Horizon 2020 programme under the Marie Skłodowska-Curie scheme (grant agreement no. 887220).

References

Abadir, K. M. and P. Paruolo (1997). Two mixed Normal densities from cointegration analysis. *Econometrica*, 671–680.

Abraham, R., J. E. Marsden, and T. Ratiu (2012). *Manifolds, tensor analysis, and applications*, Volume 75. Springer Science & Business Media.

Aldasoro, I. and I. Alves (2018). Multiplex interbank networks and systemic importance: An application to European data. *Journal of Financial Stability 35*, 17–37.

Anacleto, O. and C. Queen (2017). Dynamic chain graph models for time series network data. *Bayesian Analysis 12*(2), 491–509.

Andrade, J. A. A. and P. N. Rathie (2015). On exact posterior distributions using H-functions. *Journal of Computational and Applied Mathematics 290*, 459–475.

Andrade, J. A. A. and P. N. Rathie (2017). Exact posterior computation in non-conjugate Gaussian location-scale parameters models. *Communications in Nonlinear Science and Numerical Simulation 53*, 111–129.

Aris, R. (2012). *Vectors, tensors and the basic equations of fluid mechanics*. Courier Corporation.

Balazsi, L., L. Matyas, and T. Wansbeek (2015). The estimation of multidimensional fixed effects panel data models. *Econometric Reviews*, 1–23.

Behera, R., A. K. Nandi, and J. K. Sahoo (2020). Further results on the Drazin inverse of even order tensors. *Numerical Linear Algebra with Applications 27*(5), e2317.

Bikker, J. A. (2010). *The gravity model in international trade: Advances and applications*, Chapter An extended gravity model with substitution applied to international trade, pp. 135–164. Cambridge University Press.

Canova, F. and M. Ciccarelli (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *Journal of Econometrics 120*(2), 327–359.

Cichocki, A. (2014). Era of big data processing: A new approach via tensor networks and tensor decompositions. In *Proceedings of the International Workshop on Smart Info-Media Systems in Asia (SISA2013).*

Dawid, A. P. and S. L. Lauritzen (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 1272–1317.

De Paula, A. (2017). Econometrics of network models. In *Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress*, pp. 268–323. Cambridge University Press Cambridge.

Dobra, A. (2015). Graphical modeling of spatial health data. In *Handbook of Spatial Epidemiology*, pp. 575–594. Chapman & Hall /CRC.

Eaton, J. and S. Kortum (2002). Technology, geography, and trade. *Econometrica 70*(5), 1741–1779.

Fieler, A. C. (2011). Nonhomotheticity and bilateral trade: Evidence and a quantitative explanation. *Econometrica 79*(4), 1069–1101.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–373.

Goldsmith, J., L. Huang, and C. M. Crainiceanu (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics 23*(1), 46–64.

Guha, S. and R. Guhaniyogi (2021). Bayesian generalized sparse symmetric tensor-on-vector regression. *Technometrics 63*(2), 160–170.

Guha, S. and A. Rodriguez (2020). Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association*, 1–13.

Guhaniyogi, R. (2020). Bayesian methods for tensor regression. *Wiley StatsRef: Statistics Reference Online*, 1–18.

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017). Bayesian tensor regression. *Journal of Machine Learning Research 18*(79), 1–31.

Guhaniyogi, R. and D. Spencer (2021). Bayesian tensor response regression with an application to brain activation studies. *Bayesian Analysis 16*(4), 1221–1249.

Hackbusch, W. (2012). *Tensor spaces and numerical tensor calculus*. Springer Science & Business Media.

Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis 6*(2), 179–196.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics 9*(3), 1169–1193.

Holme, P. and J. Saramäki (2012). Temporal networks. *Physics Reports 519*(3), 97–125.

Ishwaran, H. and J. S. Rao (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics 33*(2), 730–773.

Ji, J. and Y. Wei (2018). The Drazin inverse of an even-order tensor and its application to singular tensor equations. *Computers & Mathematics with Applications 75*(9), 3402–3413.

Kapetanios, G., L. Serlenga, and Y. Shin (2021). Estimation and inference for multi-dimensional heterogeneous panel datasets with hierarchical multi-factor error structure. *Journal of Econometrics 220*(2), 504–531.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*(3), 455–500.

Koop, G., M. H. Pesaran, and S. M. Potter (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics 74*(1), 119–147.

Kostakos, V. (2009). Temporal graphs. *Physica A: Statistical Mechanics and its Applications 388*(6), 1007–1023.

Lee, N. and A. Cichocki (2018). Fundamental tensor operations for large-scale data analysis in tensor train formats. *Multidimensional Systems and Signal Processing 29*(3), 921–960.

Li, L. and X. Zhang (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association 112*(519), 1131–1146.

Li, X., D. Xu, H. Zhou, and L. Li (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences 10*(3), 520–545.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics 27*(3), 638–647.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pp. 113–162. Chapman & Hall /CRC.

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

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Pesaran, H. H. and Y. Shin (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters 58*(1), 17–29.

Rabusseau, G. and H. Kadri (2016). Low-rank regression with tensor responses. *Advances in Neural Information Processing Systems 29*, 1867–1875.

Raskutti, G., M. Yuan, H. Chen, et al. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics 47*(3), 1554–1584.

Spencer, D., R. Guhaniyogi, and R. Prado (2020). Joint Bayesian estimation of voxel activation and inter-regional connectivity in fMRI experiments. *Psychometrika 85*(4), 845–869.

Sun, W. W. and L. Li (2017). Store: Sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research 18*(1), 4908–4944.

Wang, B., H. Du, and H. Ma (2020). Perturbation bounds for DMP and CMP inverses of tensors via Einstein product. *Computational and Applied Mathematics 39*(1), 28.

Wang, H. and M. West (2009). Bayesian analysis of matrix normal graphical models. *Biometrika 96*(4), 821–834.

Xu, T., Z. Yin, T. Siliang, S. Jian, W. Fei, and Z. Yueting (2013). Logistic tensor regression for classification. In *Intelligent science and intelligent data engineering*, pp. 573–581. Springer.

Yang, Y. and D. B. Dunson (2016). Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association 111*(514), 656–669.

Yuan, M. and C.-H. Zhang (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics 16*(4), 1031–1068.

Zhang, Y. D., B. P. Naughton, H. D. Bondell, and B. J. Reich (2020). Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior. *Journal of the American Statistical Association*, 1–13.

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

Zhou, J., A. Bhattacharya, A. H. Herring, and D. B. Dunson (2015). Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association 110*(512), 1562–1576.

Supporting materials for "Bayesian Dynamic Tensor Regression"

Monica Billio
Department of Economics, Ca' Foscari University of Venice
and
Roberto Casarin
Department of Economics, Ca' Foscari University of Venice
and
Matteo Iacopini
Department of Econometrics and Data Science, Vrije Universiteit Amsterdam
Tinbergen Institute
and
Sylvia Kaufmann
Study Center Gerzensee, Foundation of the Swiss National Bank

January 19, 2022

Supplementary material

This appendix contains background results on tensors in Section S.1, and the derivation of the tensor forecast error variance decomposition in Section S.2. An example of MAR is given in Section S.3 and proofs of the remarks in Section 2 of the main paper are provided in Section S.4. Details on the prior on tensor entries are given in Section S.5. Also, Section S.6 reports the details on posterior computation and Section S.7 describes the initialisation of the inferential algorithm. A summary of simulation results is provided in Section S.8. The data used in the empirical application is described in Section S.9, and further plots of the estimation results are given in Section S.10.

S.1 Background Material on Tensor Calculus

This appendix provides the main tools used in the paper. See the supplement for further results and details. A N-order tensor is an element of the tensor product of N vector spaces. Since there exists a isomorphism between two vector spaces of dimensions N and $M < N$, it is possible to define a one-to-one map between their elements, that is, between a N-order tensor and a M-order tensor.

Definition S.1.1 (Tensor reshaping)**.** *Let $V_1, \ldots, V_N$ and $U_1, \ldots, U_M$ be vector subspaces $V_n, U_m \subseteq \mathbb{R}$ and $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_N} = V_1 \otimes \ldots \otimes V_N$ be a N-order real tensor of dimensions $I_1, \ldots, I_N$. Let $(\mathbf{v}_1, \ldots, \mathbf{v}_N)$ be a canonical basis of $\mathbb{R}^{I_1 \times \ldots \times I_N}$ and let Π_S be the projection defined as*

$$\begin{aligned} \Pi_S : & V_1 \otimes \ldots \otimes V_N \to V_{s_1} \otimes \ldots \otimes V_{s_k} \\ & \mathbf{v}_1 \otimes \ldots \otimes \mathbf{v}_N \mapsto \mathbf{v}_{s_1} \otimes \ldots \otimes \mathbf{v}_{s_k} \end{aligned}$$

with $S = \{s_1, \ldots, s_k\} \subset \{1, \ldots, N\}$. Let $(S_1, \ldots, S_M)$ be a partition of $\{1, \ldots, N\}$. The $(S_1, \ldots, S_M)$ tensor reshaping of $\mathcal{X}$ is defined as $\mathcal{X}_{(S_1, \ldots, S_M)} = (\Pi_{S_1} \mathcal{X}) \otimes \ldots \otimes (\Pi_{S_M} \mathcal{X}) = U_1 \otimes \ldots \otimes U_M$. The mapping is an isomorphism between $V_1 \otimes \ldots \otimes V_N$ and $U_1 \otimes \ldots \otimes U_M$.

The matricization is a particular case of reshaping a N-order tensor into a 2-order tensor, by choosing a mapping between the tensor modes and the rows and columns of the resulting matrix, then permuting the tensor and reshaping it, accordingly.

Definition S.1.2 (Matricization)**.** *Let $\mathcal{X}$ be a N-order tensor with dimensions $I_1, \ldots, I_N$. Let the ordered sets $\mathscr{R} = \{r_1, \ldots, r_L\}$ and $\mathscr{C} = \{c_1, \ldots, c_M\}$ be a partition of $\mathbf{N} = \{1, \ldots, N\}$. The matricized tensor is defined by*

$$\operatorname{mat}_{\mathscr{R},\mathscr{C}}(\mathcal{X}) = \mathbf{X}_{(\mathscr{R},\mathscr{C})} \in \mathbb{R}^{J \times K}, \qquad J = \prod_{n \in \mathscr{R}} I_n, \quad K = \prod_{n \in \mathscr{C}} I_n \,.$$

Indices of $\mathscr{R}, \mathscr{C}$ are mapped to the rows and the columns, respectively, and

$$\left(\mathbf{X}_{(\mathscr{R} \times \mathscr{C})}\right)_{j,k} = \mathcal{X}_{i_1, i_2, \ldots, i_N}, \quad j = 1 + \sum_{l=1}^{L} \Big((i_{r_l} - 1) \prod_{l'=1}^{l-1} I_{r'_l} \Big), \quad k = 1 + \sum_{m=1}^{M} \Big((i_{c_m} - 1) \prod_{m'=1}^{m-1} I_{c'_m} \Big).$$

The *inner product* between two $(I_1 \times \ldots \times I_N)$-dimensional tensors $\mathcal{X}, \mathcal{Y}$ is defined as

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \ldots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1,\ldots,i_N} \mathcal{Y}_{i_1,\ldots,i_N}$$

The PARAFAC(R) decomposition (e.g., see Kolda and Bader, 2009), is rank-R decomposition which represents a tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ as a finite sum of R rank-1 tensors obtained as the outer products of N vectors (called marginals) $\boldsymbol{\beta}_j^{(r)} \in \mathbb{R}^{I_j}$

$$\mathcal{B} = \sum_{r=1}^{R} \mathcal{B}_r = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_J^{(r)}.$$

Lemma S.1.1 (Contracted product – some properties)**.** *Let $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ and $\mathcal{Y} \in \mathbb{R}^{J_1 \times \ldots \times J_N \times J_{N+1} \times \ldots \times J_{N+P}}$. Let $(\mathscr{S}_1, \mathscr{S}_2)$ be a partition of $\{1, \ldots, N+P\}$, where $\mathscr{S}_1 = \{1, \ldots, N\}$, $\mathscr{S}_2 = \{N+1, \ldots, N+P\}$. It holds:*

(i) if $P = 0$ and $I_n = J_n$, $n = 1, \ldots, N$, then $\mathcal{X} \bar{\times}_N \mathcal{Y} = \langle \mathcal{X}, \mathcal{Y} \rangle = \mathrm{vec}\left(\mathcal{X}\right)' \cdot \mathrm{vec}\left(\mathcal{Y}\right)$.

(ii) if $P > 0$ and $I_n = J_n$ for $n = 1, \ldots, N$, then

$$\begin{aligned} \mathcal{X} \bar{\times}_N \mathcal{Y} &= \mathrm{vec}\left(\mathcal{X}\right) \times_1 \mathcal{Y}_{(\mathscr{S}_1, \mathscr{S}_2)} \quad \in \mathbb{R}^{j_1 \times \ldots \times j_P} \\ \mathcal{Y} \bar{\times}_N \mathcal{X} &= \mathcal{Y}_{(\mathscr{S}_1, \mathscr{S}_2)} \times_1 \mathrm{vec}\left(\mathcal{X}\right) \quad \in \mathbb{R}^{j_1 \times \ldots \times j_P}. \end{aligned}$$

(iii) let $\mathscr{R} = \{1, \ldots, N\}$ and $\mathscr{C} = \{N+1, \ldots, 2N\}$. If $P = N$ and $I_n = J_n = J_{N+n}$, $n = 1, \ldots, N$, then

$$\mathcal{X} \bar{\times}_N \mathcal{Y} \bar{\times}_N \mathcal{X} = \mathrm{vec}\left(\mathcal{X}\right)' \mathbf{Y}_{(\mathscr{R}, \mathscr{C})} \, \mathrm{vec}\left(\mathcal{X}\right).$$

(iv) let $M = N + P$, then $\mathcal{X} \circ \mathcal{Y} = \underline{\mathcal{X}} \bar{\times}_1 \underline{\mathcal{Y}}^T$, where $\underline{\mathcal{X}}, \underline{\mathcal{Y}}$ are $(I_1 \times \ldots \times I_N \times 1)$- and $(J_1 \times \ldots \times J_M \times 1)$-dimensional tensors, respectively, given by $\underline{\mathcal{X}}_{:,\ldots,:,1} = \mathcal{X}$, $\underline{\mathcal{Y}}_{:,\ldots,:,1} = \mathcal{Y}$ and $\underline{\mathcal{Y}}^T_{j_1,\ldots,j_M,j_{M+1}} = \underline{\mathcal{Y}}_{j_{M+1},j_M,\ldots,j_1}$.

Proof. Case (i). By definition of contracted product and tensor scalar product

$$\mathcal{X} \bar{\times}_N \mathcal{Y} = \sum_{i_1=1}^{I_1} \ldots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1,\ldots,i_N} \mathcal{Y}_{i_1,\ldots,i_N} = \langle \mathcal{X}, \mathcal{Y} \rangle = \mathrm{vec}\left(\mathcal{X}\right)' \cdot \mathrm{vec}\left(\mathcal{Y}\right).$$

Case (ii). Define $I^* = \prod_{n=1}^{N} I_n$ and $k = 1 + \sum_{j=1}^{N}(i_j - 1)\prod_{m=1}^{j-1} I_m$. By definition of contracted product and tensor scalar product

$$\mathcal{X}\bar{\times}_N\mathcal{Y} = \sum_{i_1=1}^{I_1}\dots\sum_{i_N=1}^{I_N}\mathcal{X}_{i_1,\dots,i_N}\mathcal{Y}_{i_1,\dots,i_N,j_{N+1},\dots,j_{N+P}} = \sum_{k=1}^{I^*}\mathcal{X}_k\mathcal{Y}_{k,j_{N+1},\dots,j_{N+P}}\,.$$

Note that the one-to-one correspondence established by the mapping between k and $(i_1,\dots,i_N)$ corresponds to that of the vectorization of a $(I_1\times\dots\times I_N)$-dimensional tensor. It also corresponds to the mapping established by the tensor reshaping of a $(N+P)$-order tensor with dimensions $I_1,\dots,I_N,J_{N+1},\dots,J_{N+P}$ into a $(P+1)$-order tensor with dimensions $I^*,J_{N+1},\dots,J_{N+P}$. Let $\mathscr{S}_1 = \{1,\dots,N\}$, then

$$\mathcal{X}\bar{\times}_N\mathcal{Y} = \sum_{i_1=1}^{I_1}\dots\sum_{i_N=1}^{I_N}\mathcal{X}_{i_1,\dots,i_N}\mathcal{Y}_{i_1,\dots,i_N,:,\dots,:} = \sum_{s_1=1}^{|\mathscr{S}_1|}\mathbf{x}_{s_1}\bar{\mathcal{Y}}_{s_1,:,\dots,:}$$

where $\bar{\mathcal{Y}} = \mathrm{reshape}_{(\mathscr{S}_1,N+1,\dots,N+P)}(\mathcal{Y})$. Following the same approach, and defining $\mathscr{S}_2 = \{N+1,\dots,N+P\}$, we obtain the second part of the result.

Case (iii). We follow the same strategy adopted in case b). Let $\mathbf{x} = \mathrm{vec}\big(\mathcal{X}\big)$, $S_1 = \{1,\dots,N\}$ and $S_2 = \{N+1,\dots,N+P\}$, such that $(S-1,S_2)$ is a partition of $\{1,\dots,N+P\}$. Let k,k' be defined as in case b). Then

$$\mathcal{X}\bar{\times}_N\mathcal{Y}\bar{\times}_N\mathcal{X} = \sum_{i_1=1}^{I_1}\dots\sum_{i_N=1}^{I_N}\sum_{i'_1=1}^{I_1}\dots\sum_{i'_N=1}^{I_N}\mathcal{X}_{i_1,\dots,i_N}\mathcal{Y}_{i_1,\dots,i_N,i'_1,\dots,i'_N}\mathcal{X}_{i'_1,\dots,i'_N}$$

$$= \sum_{k=1}^{I^*}\sum_{i'_1=1}^{I_1}\dots\sum_{i'_N=1}^{I_N}\mathbf{x}_k\mathcal{Y}_{k,i'_1,\dots,i'_N}\mathcal{X}_{i'_1,\dots,i'_N} = \sum_{k=1}^{I^*}\sum_{k'=1}^{I^*}\mathbf{x}_k\mathcal{Y}_{k,k'}\mathbf{x}_{k'} = \mathrm{vec}\big(\mathcal{X}\big)'\mathcal{Y}_{(S_1,S_2)}\,\mathrm{vec}\big(\mathcal{X}\big).$$

Case (iv). Let $\mathbf{i} = (i_1,\dots,i_N)$ and $\mathbf{j} = (j_1,\dots,j_M)$ be two multi-indexes. By the definition of outer and contracted product we get $(\mathcal{X}\circ\mathcal{Y})_{\mathbf{i},\mathbf{j}} = \underline{\mathcal{X}}_{\mathbf{i},1}\underline{\mathcal{Y}}_{1,\mathbf{j}} = (\underline{\mathcal{X}}\bar{\times}_1\underline{\mathcal{Y}}^T)_{\mathbf{i},\mathbf{j}}$. Therefore, with a slight abuse of notation, we use $\underline{\mathcal{Y}} = \mathcal{Y}$ and write $\mathcal{Y}\circ\mathcal{Y} = \mathcal{Y}\bar{\times}_1\mathcal{Y}^T$, when the meaning of the products is clear form the context. $\square$

Lemma S.1.2 (Kronecker - matricization)**.** *Let X_n be a $I_n\times I_n$ matrix, for $n = 1,\dots,N$, and let $\mathcal{X} = X_1\circ\dots\circ X_N$ be the $(I_1\times\dots\times I_N\times I_1\times\dots\times I_N)$-dimensional tensor obtained as the outer product of the matrices $X_1,\dots,X_N$. Let $(\mathscr{S}_1,\mathscr{S}_2)$ be a partition of $I_{\mathbf{N}} = \{1,\dots,2N\}$, where $\mathscr{S}_1 = \{1,\dots,N\}$ and $\mathscr{S}_2 = \{N+1,\dots,N\}$. Then $\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)} = \mathbf{X}_{(\mathscr{R},\mathscr{C})} = (X_N\otimes\dots\otimes X_1)$.*

Proof. Use the pair of indices (i_n, i'_n) for the entries of the matrix X_n, $n = 1, \dots, N$. By definition of outer product $(X_1 \circ \dots \circ X_N)_{i_1,\dots,i_N,i'_1,\dots,i'_N} = (X_1)_{i_1,i'_1} \cdot \dots \cdot (X_N)_{i_N,i'_N}$. By definition of matricization, $\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)} = \mathbf{X}_{(\mathscr{R},\mathscr{C})}$. Moreover $(\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)})_{h,k} = \mathcal{X}_{i_1,\dots,i_{2N}}$ with $h = \sum_{p=1}^{N}(i_{S_{1,p}} - 1)\prod_{q=1}^{p-1} J_{S_{1,p}}$ and $k = \sum_{p=1}^{N}(i_{S_{2,p}} - 1)\prod_{q=1}^{p-1} J_{S_{2,p}}$. By definition of the Kronecker product, the entry (h', k') of $(X_N \otimes \dots \otimes X_1)$ is $(X_N \otimes \dots \otimes X_1)_{h',k'} = (X_N)_{i'_N,i'_N} \cdot \dots \cdot (X_1)_{i_1,i'_1}$, where $h' = \sum_{p=1}^{N}(i_{S_{1,p}} - 1)\prod_{q=1}^{p-1} J_{S_{1,p}}$ and $k' = \sum_{p=1}^{N}(i_{S_{2,p}} - 1)\prod_{q=1}^{p-1} J_{S_{2,p}}$. Since $h = h'$ and $k = k'$ and the associated elements of $\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)}$ and $(X_N \otimes \dots \otimes X_1)$ are the same, the result follows. $\square$

Lemma S.1.3 (Outer product and vectorization)**.** *Let $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n$ be vectors such that $\boldsymbol{\alpha}_i$ has length d_i, for $i = 1, \dots, n$. Then, for each $j = 1, \dots, n$, it holds*

$$\operatorname{vec}\big(\mathop{\circ}_{i=1}^{n} \boldsymbol{\alpha}_i\big) = \mathop{\otimes}_{i=1}^{n} \boldsymbol{\alpha}_{n-i+1} = \big(\boldsymbol{\alpha}_n \otimes \dots \otimes \boldsymbol{\alpha}_{j+1} \otimes \mathbf{I}_{d_j} \otimes \boldsymbol{\alpha}_{j-1} \otimes \dots \otimes \boldsymbol{\alpha}_1\big)\boldsymbol{\alpha}_j.$$

Proof. The result follows from the definitions of vectorisation operator and outer product. For $n = 2$, the result follows directly from

$$\operatorname{vec}\big(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2\big) = \operatorname{vec}\big(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}'_2\big) = \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_1 = (\boldsymbol{\alpha}_2 \otimes \mathbf{I}_{d_1})\boldsymbol{\alpha}_1 = (\mathbf{I}_{d_2} \otimes \boldsymbol{\alpha}_1)\boldsymbol{\alpha}_2.$$

For $n > 2$ consider, without loss of generality, $n = 3$ (an analogous proof holds for $n > 3$). Then, from the definitions of outer product and Kronecker product we have

$$\begin{aligned}
&\operatorname{vec}\big(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2 \circ \boldsymbol{\alpha}_3\big) = \\
&= (\boldsymbol{\alpha}'_1 \cdot \alpha_{2,1}\alpha_{3,1}, \dots, \boldsymbol{\alpha}'_1 \cdot \alpha_{2,d_2}\alpha_{3,1}, \boldsymbol{\alpha}'_1 \cdot \alpha_{2,1}\alpha_{3,2}, \dots, \boldsymbol{\alpha}'_1 \cdot \alpha_{2,d_2}\alpha_{3,2}, \dots, \boldsymbol{\alpha}'_1 \cdot \alpha_{2,d_2}\alpha_{3,d_3})' \\
&= \boldsymbol{\alpha}_3 \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_1 = (\boldsymbol{\alpha}_3 \otimes \boldsymbol{\alpha}_2 \otimes \mathbf{I}_{d_1})\boldsymbol{\alpha}_1 = (\boldsymbol{\alpha}_3 \otimes \mathbf{I}_{d_2} \otimes \boldsymbol{\alpha}_1)\boldsymbol{\alpha}_2 = (\mathbf{I}_{d_3} \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_1)\boldsymbol{\alpha}_3.
\end{aligned}$$

$\square$

Let $\mathcal{X}, \mathcal{Y}$ be two $(I_1 \times \dots \times I_N)$-dimensional tensors. The *Hadamard product* between them, $\mathcal{Z} = \mathcal{X} \odot \mathcal{Y}$, is the $(I_1 \times \dots \times I_N)$-dimensional tensor $\mathcal{Z}$ defined by the element-wise multiplication

$$\mathcal{Z}_{i_1,\dots,i_N} = (\mathcal{X} \odot \mathcal{Y})_{i_1,\dots,i_N} = \mathcal{X}_{i_1,\dots,i_N} \mathcal{Y}_{i_1,\dots,i_N}.$$

We introduce two multilinear operators acting on tensors (see Kolda, 2006, for further details).

Definition S.1.3 (Tucker operator)**.** *Let* $\mathcal{Y} \in \mathbb{R}^{J_1 \times \ldots \times J_N}$ *and* $\mathbf{N} = \{1, \ldots, N\}$*. Let* $(A_n)_n$ *be a collection of* N *matrices such that* $A_n \in \mathbb{R}^{I_n \times J_n}$*. The Tucker operator is defined as*

$$[\![\mathcal{Y}; A_1, \ldots, A_N]\!] = \mathcal{Y} \bar{\times}_1 A_1 \bar{\times}_1 A_2 \ldots \bar{\times}_1 A_N,$$

54 *and the resulting tensor has size* $I_1 \times \ldots \times I_N$*.*

We now define some useful tensor decompositions. The Tucker decomposition is a higher-order generalization of the Principal Component Analysis (PCA): a tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ is decomposed into the product (along the corresponding modes) of a "core" tensor $\mathcal{G} \in \mathbb{R}^{g_1 \times \ldots \times g_N}$ and factor matrices $A^{(m)} \in \mathbb{R}^{I_m \times J_m}$, $m = 1, \ldots, N$

$$\mathcal{B} = \mathcal{G} \bar{\times}_1 A^{(1)} \bar{\times}_1 \ldots \bar{\times}_1 A^{(N)} = \sum_{i_1=1}^{g_1} \ldots \sum_{i_N=1}^{g_N} \mathcal{G}_{i_1,\ldots,i_N} \mathbf{a}_{i_1}^{(1)} \circ \ldots \circ \mathbf{a}_{i_N}^{(N)} \tag{S1}$$

where $\mathbf{a}_{i_l}^{(m)} \in \mathbb{R}^{g_m}$ is the m-th column of the matrix $A^{(m)}$. As a result, each entry of the tensor is obtained as

$$\mathcal{B}_{j_1,\ldots,j_N} = \sum_{i_1=1}^{g_1} \cdots \sum_{i_N=1}^{g_N} \mathcal{G}_{i_1,\ldots,i_N} \cdot A_{i_1,j_1}^{(1)} \cdots A_{i_N,j_N}^{(N)} \tag{S2}$$

The PARAFAC(R) decomposition[1], is rank-R decomposition which represents a tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ as a finite sum of R rank-1 tensors obtained as the outer products of N vectors (called marginals) $\boldsymbol{\beta}_j^{(r)} \in \mathbb{R}^{I_j}$, $j = 1, \ldots, J$

$$\mathcal{B} = \sum_{r=1}^{R} \mathcal{B}_r = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_J^{(r)}. \tag{S3}$$

55 Fig. 1 provides a graphical representation of this decomposition for a 3-order tensor.

Definition S.1.4 (Kruskal operator)**.** *Let* $\mathbf{N} = \{1, \ldots, N\}$ *and* $(A_n)_n$ *be a collection of* N *matrices such that* $A_n \in \mathbb{R}^{I_n \times R}$ *for* $n \in \mathbf{N}$*. Let* $\mathcal{I}$ *be the identity tensor of size* $R \times \ldots \times R$*, i.e. a tensor having ones along the superdiagonal and zeros elsewhere. The Kruskal operator is defined as*

$$\mathcal{X} = [\![A_1, \ldots, A_N]\!] = [\![\mathcal{I}; A_1, \ldots, A_N]\!],$$

[1]See Harshman (1970). Some authors (e.g. Carroll and Chang, 1970; Kiers, 2000) use the term CODECOMP or CP instead of PARAFAC.

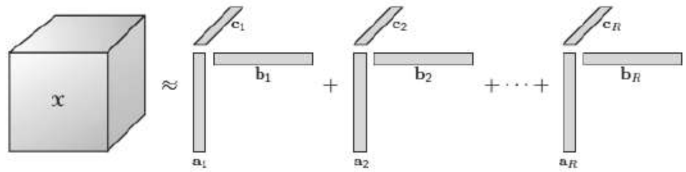

Figure 1: PARAFAC decomposition of $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, with $\mathbf{a}_r \in \mathbb{R}^{I_1}$, $\mathbf{b}_r \in \mathbb{R}^{I_2}$ and $\mathbf{c}_r \in \mathbb{R}^{I_3}$, $r = 1, \ldots, R$. Figure from Kolda and Bader (2009).

with $\mathcal{X}$ a tensor of size $I_1 \times \ldots \times I_N$. An alternative representation is obtained by defining $\mathbf{a}_n^{(r)}$ the r-th column of the matrix A_n and using the outer product

$$\mathcal{X} = [\![A_1, \ldots, A_N]\!] = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \circ \ldots \circ \mathbf{a}_N^{(r)}.$$

By exploiting the Khatri-Rao product $\odot^K$ (i.e. the column-wise Kronecker product for $A \in \mathbb{R}^{I \times K}$, $B \in \mathbb{R}^{J \times K}$ defined as $A \odot^K B = (\mathbf{a}_{:,1} \otimes \mathbf{b}_{:,1}, \ldots, \mathbf{a}_{:,K} \otimes \mathbf{b}_{:,K})$) in combination with the mode-n matricization and the vectorization operators, we get the following additional representations of $\mathcal{X} = [\![A_1, \ldots, A_N]\!]$

$$\begin{aligned} \mathbf{X}_{(n)} &= A_n (A_N \odot^K \ldots \odot^K A_{n+1} \odot^K A_{n-1} \odot^K \ldots \odot^K A_1)' \\ \mathrm{vec}\left(\mathcal{X}\right) &= (A_N \odot^K \ldots \odot^K A_1) \mathbf{1}_R \end{aligned}$$

56 *where $\mathbf{1}_R$ is a vector of ones of length R.*

Remark S.1.1. *Let $\mathcal{X}$ be a N-order tensor of dimensions $I_1 \times \ldots \times I_N$ and let $I^* = \prod_{i=1}^N I_i$. Then there exists a $I^* \times I^*$ vec-permutation (or commutation) matrix $K_{1 \to n}$ such that*

$$K_{1\to n} \, \mathrm{vec}\left(\mathcal{X}\right) = K_{1\to n} \, \mathrm{vec}\left(\mathbf{X}_{(1)}\right) = \mathrm{vec}\left(\mathbf{X}_{(n)}\right) = \mathrm{vec}\left(\mathbf{X}_{(1)}^{T_\sigma}\right) = \mathrm{vec}\left(\mathcal{X}^{T_\sigma}\right),$$

where $\mathbf{X}_{(1)}^{T_\sigma} = \left(\mathcal{X}^{T_\sigma}\right)_{(1)} = \mathbf{X}_{(n)}$ is the mode-1 matricization of the transposed tensor $\mathcal{X}^{T_\sigma}$ according to the permutation σ which exchanges modes 1 and n, leaving the others unchanged. That is, for $i_j \in \{1, \ldots, I_j\}$ and $j = 1, \ldots, N$

$$\sigma(i_j) = \begin{cases} 1 & j = n \\ n & j = 1 \\ i_j & j \neq 1, n \end{cases}$$

Lemma S.1.4 (Tensor – matrix Normal)**.** *Let $\mathcal{X}$ be a N-order random tensor with dimensions $I_1,\ldots,I_N$ and let $\mathbf{N}=\{1,\ldots,N\}$ be partitioned by the index sets $\mathscr{R}=\{r_1,\ldots,r_m\}\subset\mathbf{D}$ and $\mathscr{C}=\{c_1,\ldots,c_p\}\subset\mathbf{N}$, i.e. $\mathbf{N}=\mathscr{R}\cup\mathscr{C}$, $\mathscr{R}\cap\mathscr{C}=\emptyset$ and $N=m+p$. Then*

$$\mathcal{X}\sim\mathcal{N}_{I_1,\ldots,I_N}(\mathcal{M},\Sigma_1,\ldots,\Sigma_N)\iff\mathbf{X}_{(\mathscr{R}\times\mathscr{C})}\sim\mathcal{N}_{m,p}(\mathbf{M}_{(\mathscr{R}\times\mathscr{C})},\Sigma_1,\Sigma_2),$$

57 *with $\Sigma_1=\Sigma_{r_m}\otimes\ldots\otimes\Sigma_{r_1}$ and $\Sigma_2=\Sigma_{c_p}\otimes\ldots\otimes\Sigma_{c_1}$.*

Proof. We demonstrate the statement for $\mathscr{R}=\{n\}$, $n\in\mathbf{N}$, however the results follows from the same steps also in the general case $\#\mathscr{R}>1$. The strategy is to demonstrate that the probability density functions of the two distributions coincide. To this aim consider separately the exponent and the normalizing constant. Define $I_{-j}=\prod_{i=1,n\neq j}^N I_i$ and $I_{\mathbf{N}}=\{I_1,\ldots,I_N\}$, then for the normalizing constant we have

$$(2\pi)^{-\frac{\prod_i I_i}{2}}\left|\Sigma_1\right|^{-\frac{I_{-1}}{2}}\cdots\left|\Sigma_n\right|^{-\frac{I_{-n}}{2}}\cdots\left|\Sigma_N\right|^{-\frac{I_{-N}}{2}}= \tag{S4}$$

$$=(2\pi)^{-\frac{\prod_i I_i}{2}}\left|\Sigma_1\right|^{-\frac{I_{-1}}{2}}\cdots\left|\Sigma_{n-1}\right|^{-\frac{I_{-(n-1)}}{2}}\left|\Sigma_{n+1}\right|^{-\frac{I_{-(n+1)}}{2}}\cdots\left|\Sigma_N\right|^{-\frac{I_{-N}}{2}}\left|\Sigma_n\right|^{-\frac{I_{-n}}{2}}$$

$$=(2\pi)^{-\frac{\prod_i I_i}{2}}\left|\Sigma_N\otimes\ldots\otimes\Sigma_{n-1}\otimes\Sigma_{n+1}\otimes\ldots\otimes\Sigma_N\right|^{-\frac{n}{2}}\left|\Sigma_n\right|^{-\frac{I_{-n}}{2}}. \tag{S5}$$

Concerning the exponent, let $\mathbf{i}=(i_1,\ldots,i_N)$ and, for ease of notation, define $\mathcal{Y}=\mathcal{X}-\mathcal{M}$ and $\mathcal{U}=(\Sigma_N^{-1}\circ\ldots\circ\Sigma_1^{-1})$. By the definition of contracted and outer products, it holds

$$\mathcal{Y}\bar{\times}_N\mathcal{U}\bar{\times}_N\mathcal{Y}=\sum_{i_1,\ldots,i_n,\ldots,i_N}\sum_{i'_1,\ldots,i'_n,\ldots,i'_N}y_{i_1,\ldots,i_N}(u^{-1}_{i_1,i'_1}\cdot\ldots\cdot u_{i_n,i'_n}\cdot\ldots\cdot u^{-1}_{i_N,i'_N})y_{i'_1,\ldots,i'_n,\ldots,i'_N}. \tag{S6}$$

Define $\mathbf{j}=\sigma(\mathbf{i})$, where σ is the permutation defined in Remark S.1.1 exchanging i_1 with i_n, $n\in\{2,\ldots,N\}$. Then the previous equation can be rewritten as

$$\begin{aligned}\mathcal{Y}\bar{\times}_N\mathcal{U}\bar{\times}_N\mathcal{Y}&=\sum_{j_1,\ldots,j_N}\sum_{j'_1,\ldots,j'_N}y_{j_n,\ldots,j_1,\ldots,i_N}(u^{-1}_{j_n,j'_n}\cdots u^{-1}_{j_1,j'_1}\cdots u^{-1}_{i_N,i'_N})y_{j'_n,\ldots,j'_1,\ldots,i'_N}\\&=\mathcal{Y}^\sigma\bar{\times}_N\left(\Sigma_1^{-1}\circ\ldots\circ\Sigma_N^{-1}\right)^\sigma\bar{\times}_N\mathcal{Y}^\sigma\end{aligned}$$

where $\mathcal{Y}^\sigma$ is the transpose tensor of $\mathcal{Y}$ (see Pan, 2014) obtained by permuting the first and the n-th modes and similarly for the N-order tensor $(\Sigma_1^{-1}\circ\ldots\circ\Sigma_N^{-1})^\sigma$. Let $(\mathscr{S}_1,\mathscr{S}_2)$, with $\mathscr{S}_1=\{1,\ldots,N\}$ and $\mathscr{S}_2=\{N+1,\ldots,2N\}$, be a partition of $\{1,\ldots,2N\}$. By vectorizing eq. (S6) and exploiting the results in Theorem S.1.1 and Theorem S.1.2, we have

$$\mathcal{Y}\bar{\times}_N\mathcal{U}\bar{\times}_N\mathcal{Y}=\operatorname{vec}\left(\mathcal{Y}\right)'\cdot\mathcal{U}_{(\mathscr{S}_1,\mathscr{S}_2)}\cdot\operatorname{vec}\left(\mathcal{Y}\right) \tag{S7}$$

$$
\begin{aligned}
&= \operatorname{vec}\left(\mathcal{Y}\right)' \cdot \left(\Sigma_N^{-1} \otimes \ldots \otimes \Sigma_n^{-1} \otimes \ldots \otimes \Sigma_1^{-1}\right) \cdot \operatorname{vec}\left(\mathcal{Y}\right) \\
&= \operatorname{vec}\left(\mathcal{Y}^\sigma\right)' \cdot \left(\Sigma_N^{-1} \otimes \ldots \otimes \Sigma_1^{-1} \otimes \Sigma_n^{-1}\right) \cdot \operatorname{vec}\left(\mathcal{Y}^\sigma\right) \\
&= \operatorname{vec}\left(\mathbf{Y}_{(n)}\right)' \cdot \left(\Sigma_N^{-1} \otimes \ldots \otimes \Sigma_1^{-1} \otimes \Sigma_n^{-1}\right) \cdot \operatorname{vec}\left(\mathbf{Y}_{(n)}\right) \\
&= \operatorname{vec}\left(\mathbf{Y}_{(n)}\right)' \cdot \operatorname{vec}\left(\Sigma_n^{-1} \cdot \mathbf{Y}_{(n)} \cdot \left(\Sigma_N^{-1} \otimes \ldots \otimes \Sigma_1^{-1}\right)\right) \\
&= \operatorname{tr}\left(\mathbf{Y}_{(n)}' \cdot \Sigma_n^{-1} \cdot \mathbf{Y}_{(n)} \cdot \left(\Sigma_N^{-1} \otimes \ldots \otimes \Sigma_1^{-1}\right)\right) \\
&= \operatorname{tr}\left(\left(\Sigma_N^{-1} \otimes \ldots \otimes \Sigma_1^{-1}\right)\left(\mathbf{X}_{(n)} - \mathbf{M}_{(n)}\right)'\Sigma_n^{-1}\left(\mathbf{X}_{(n)} - \mathbf{M}_{(n)}\right)\right). \qquad \text{(S8)}
\end{aligned}
$$

Since the term in (S4) and (S7) are the normalizing constant and the exponent of the tensor normal distribution, whereas (S5) and (S8) are the corresponding expressions for the desired matrix normal distribution, the result is proved for the case $\#\mathscr{R} = 1$. In the general case $\#\mathscr{R} = r > 1$ the proof follows from the same reasoning, by substituting the permutation σ with another permutation σ' which exchanges the modes of the tensor such that the first r modes of the transpose tensor $\mathcal{Y}^{\sigma'}$ correspond to the elements of $\mathscr{R}$. $\square$

S.2 Forecast error variance decomposition

From the results in eqs. (11)-(12) of the main paper, we obtain the forecast error variance decomposition (tFEVD) for the tensor autoregressive model in each of the two cases. The tFEVD $\theta_{i,j}(h)$ measures the proportion of the h-step ahead forecast error variance of variable i that is accounted for by the innovations in variable j, in the VAR formulation of the model. Recently, Lanne and Nyberg (2016) have introduced a modification to the FEVD obtained from the GIRF of Koop et al. (1996), $\theta^*_{i,j}(h)$, which has unit sum. Denoting by $IRF(h)$ an impulse response function at horizon h, the corresponding tFEVD and its modification are, respectively,

$$
\theta_{ij}(h) = \frac{\sum_{k=0}^{h} IRF_{ij}^2(k)}{\sum_{k=0}^{h}\sum_{j=0}^{I^*} IRF_{ij}^2(k)}, \qquad \theta^*_{i,j}(h) = \frac{\sum_{k=0}^{h} (\psi_{ij}^G(k;n))^2}{\sum_{k=0}^{h}\sum_{j=0}^{I^*} (\psi_{ij}^G(k;n))^2}.
$$

The orthogonalised tensor forecast error variance decomposition (OtFEVD) by construction sums (over j) to 1. In this case $\delta_j^* = 1$, and all the other $I^* - 1$ entries are zero (equivalent to $\boldsymbol{\delta}^* = \mathbf{e}_j$). The OtFEVD is given by

$$
\theta^O_{i,j}(h) = \frac{\sum_{k=0}^{h} (\psi_{ij}^O(k;n))^2}{\sum_{k=0}^{h}\sum_{j=0}^{I^*} (\psi_{ij}^O(k;n))^2} = \frac{\sum_{k=0}^{h} (\mathbf{e}_i'\Psi_k L P \mathbf{e}_j)^2}{\sum_{k=0}^{h} \mathbf{e}_i'(\Psi_k L) D (\Psi_k L)'\mathbf{e}_i}.
$$

Consider the case $\delta^*_j = \sqrt{D_{jj}}$, with all the other $I^* - 1$ entries being zero (equivalent to $\boldsymbol{\delta}^* = \sqrt{D_{jj}}\mathbf{e}_j$). The generalised tensor forecast error variance decomposition (GtFEVD) does not sum to 1, and is

$$\theta^G_{i,j}(h) = \frac{\sum_{k=0}^{h}(\psi^G_{ij}(k;n))^2}{\sum_{k=0}^{h}\sum_{j=0}^{I^*}(\psi^G_{ij}(k;n))^2} = \frac{\sum_{k=0}^{h}(\mathbf{e}_i'\Psi_k LDD_{jj}^{-1/2}\mathbf{e}_j)^2}{\sum_{k=0}^{h}\mathbf{e}_i'(\Psi_k L)D(\Psi_k L)'\mathbf{e}_i} \tag{S9}$$

Finally, the modified tFEVD applied to the tensor GIRF (S9) yields

$$\theta^{G*}_{i,j}(h) = \frac{\sum_{k=0}^{h}(\mathbf{e}_i'\Psi_k LDD_{jj}^{-1/2}\mathbf{e}_j)^2}{\sum_{k=0}^{h}\sum_{j=1}^{I^*}\mathbf{e}_i'\Psi_k LDD_{jj}^{-1/2}\mathbf{e}_j} = \frac{\sum_{k=0}^{h}(\mathbf{e}_i'\Psi_k LDD_{jj}^{-1/2}\mathbf{e}_j)^2}{\sum_{k=0}^{h}\mathbf{e}_i'(\Psi_k L)D\Lambda D'(\Psi_k L)'\mathbf{e}_i},$$

where $\Lambda = \mathrm{diag}(D_{11}^{-1}, \ldots, D_{I^*I^*}^{-1})$.

S.3 Example: MAR(1)

To facilitate the understanding of the model in eq. (5), this section shows a special case of the general model in eq. (5), that we call the matrix autoregressive model, or MAR(p). We illustrate in a toy example the case with only the lagged dependent variable (i.e., $\mathcal{Y}_{t-1}$) as regressor. Assuming $N = 2$, $p = 1$ and $I_1 = I_2 = 2$, we obtain a matrix autoregressive model (i.e. with $\mathcal{Y}_t = Y_t$, $\mathcal{E}_t = E_t$) with one lag. Denoting $\mathrm{vec}\left(\mathcal{Y}_t\right) = \mathbf{y}_t$ and $\mathrm{vec}\left(\mathcal{E}_t\right) = \boldsymbol{\epsilon}_t$, as follows

$$\mathcal{Y}_t = \begin{pmatrix} y_{11,t} & y_{12,t} \\ y_{21,t} & y_{22,t} \end{pmatrix} \implies \mathrm{vec}\left(\mathcal{Y}_t\right) = (y_{11,t}, y_{12,t}, y_{21,t}, y_{22,t})' = (y_{1,t}, y_{2,t}, y_{3,t}, y_{4,t})'$$

$$\mathcal{B} = (\mathcal{B}_{::1}, \mathcal{B}_{::2}, \mathcal{B}_{::3}, \mathcal{B}_{::4}), \qquad \text{with } \mathcal{B}_{::k} = B_k = \begin{pmatrix} b_{11k} & b_{12k} \\ b_{21k} & b_{22k} \end{pmatrix}$$

$$\mathcal{E}_t = \begin{pmatrix} \epsilon_{11,t} & \epsilon_{12,t} \\ \epsilon_{21,t} & \epsilon_{22,t} \end{pmatrix} \implies \mathrm{vec}\left(\mathcal{E}_t\right) = (\epsilon_{11,t}, \epsilon_{12,t}, \epsilon_{21,t}, \epsilon_{22,t})' = (\epsilon_{1,t}, \epsilon_{2,t}, \epsilon_{3,t}, \epsilon_{4,t})'.$$

Therefore, model (5) becomes

$$\mathcal{Y}_t = \mathcal{B}\bar{\times}_1\mathcal{Y}_{t-1} + \mathcal{E}_t \implies Y_t = \mathcal{B}\bar{\times}_1 Y_{t-1} + E_t$$

$$\begin{pmatrix} y_{11,t} & y_{12,t} \\ y_{21,t} & y_{22,t} \end{pmatrix} = \mathcal{B}_{::1}\mathbf{y}_{1,t-1} + \ldots + \mathcal{B}_{::4}\mathbf{y}_{4,t-1} + \begin{pmatrix} \epsilon_{11,t} & \epsilon_{12,t} \\ \epsilon_{21,t} & \epsilon_{22,t} \end{pmatrix}$$

$$= \begin{pmatrix} b_{11,1} & b_{12,1} \\ b_{21,1} & b_{22,1} \end{pmatrix} \mathbf{y}_{1,t-1} + \ldots + \begin{pmatrix} b_{11,4} & b_{12,4} \\ b_{21,4} & b_{22,4} \end{pmatrix} \mathbf{y}_{4,t-1} + \begin{pmatrix} \epsilon_{11,t} & \epsilon_{12,t} \\ \epsilon_{21,t} & \epsilon_{22,t} \end{pmatrix}.$$

Assuming a PARAFAC(R) decomposition on the tensor coefficient $\mathcal{B}$ yields

$$\begin{aligned}
\mathcal{B} &= \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} = \sum_{r=1}^{R} \begin{pmatrix} \beta_{1,1}^{(r)} \\ \beta_{1,2}^{(r)} \end{pmatrix} \circ \begin{pmatrix} \beta_{2,1}^{(r)} \\ \beta_{2,2}^{(r)} \end{pmatrix} \circ \begin{pmatrix} \beta_{3,1}^{(r)} \\ \beta_{3,2}^{(r)} \\ \beta_{3,3}^{(r)} \\ \beta_{3,4}^{(r)} \end{pmatrix} \\
&= \sum_{r=1}^{R} \begin{pmatrix} \beta_{1,1}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,1}^{(r)}\beta_{2,2}^{(r)} \\ \beta_{1,2}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,2}^{(r)}\beta_{2,2}^{(r)} \end{pmatrix} \circ \begin{pmatrix} \beta_{3,1}^{(r)} \\ \beta_{3,2}^{(r)} \\ \beta_{3,3}^{(r)} \\ \beta_{3,4}^{(r)} \end{pmatrix} \\
&= \left(\sum_{r=1}^{R} \beta_{3,1}^{(r)} \begin{pmatrix} \beta_{1,1}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,1}^{(r)}\beta_{2,2}^{(r)} \\ \beta_{1,2}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,2}^{(r)}\beta_{2,2}^{(r)} \end{pmatrix}, \; \ldots, \; \sum_{r=1}^{R} \beta_{3,4}^{(r)} \begin{pmatrix} \beta_{1,1}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,1}^{(r)}\beta_{2,2}^{(r)} \\ \beta_{1,2}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,2}^{(r)}\beta_{2,2}^{(r)} \end{pmatrix} \right) \\
&= (\mathcal{B}_{::1}, \mathcal{B}_{::2}, \mathcal{B}_{::3}, \mathcal{B}_{::4}),
\end{aligned}$$

where, for each $i = 1, \ldots, 4$, we have

$$\mathcal{B}_{::k} = \sum_{r=1}^{R} \beta_{3,k}^{(r)} \begin{pmatrix} \beta_{1,1}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,1}^{(r)}\beta_{2,2}^{(r)} \\ \beta_{1,2}^{(r)}\beta_{2,1}^{(r)} & \beta_{1,2}^{(r)}\beta_{2,2}^{(r)} \end{pmatrix} = \begin{pmatrix} b_{11k} & b_{12k} \\ b_{21k} & b_{22k} \end{pmatrix},$$

hence, by choosing a PARAFAC(R) decomposition, we are assuming

$$b_{ijk} = \sum_{r=1}^{R} \beta_{1,i}^{(r)} \beta_{2,j}^{(r)} \beta_{3,k}^{(r)}, \qquad i = 1,2, \; j = 1,2, \; k = 1, \ldots, 4.$$

S.4 Proofs of the results in the main paper

In this section we provide the derivation of the results in the main paper. We start by recalling a relationship between between the outer product, the Kronecker product and the ordinary matrix product. For two vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^m$ it holds $\mathbf{u} \otimes \mathbf{v}' = \mathbf{u} \circ \mathbf{v} = \mathbf{u}\mathbf{v}'$.

Proof of Proposition 2.1. Denote with L the lag operator, s.t. $L\mathcal{Y}_t = \mathcal{Y}_{t-1}$, by properties of the contracted product in Theorem S.1.1, case (iv), we get $(\mathcal{I} - \widetilde{\mathcal{A}}_1 L) \bar{\times}_N \mathcal{Y}_t = \widetilde{\mathcal{A}}_0 +$

$\widetilde{\mathcal{B}}\bar{\times}_M\mathcal{X}_t+\mathcal{E}_t$. We apply to both sides the operator $(\mathcal{I}+\widetilde{\mathcal{A}}_1 L+\widetilde{\mathcal{A}}_1^2 L^2+\ldots+\widetilde{\mathcal{A}}_1^{t-1}L^{t-1})$, take $t\to\infty$, and get

$$\lim_{t\to\infty}(\mathcal{I}-\widetilde{\mathcal{A}}_1^t L^t)\bar{\times}_N\mathcal{Y}_t=\Big(\sum_{k=0}^{\infty}\widetilde{\mathcal{A}}_1^k L^k\Big)\bar{\times}_N(\widetilde{\mathcal{A}}_0+\widetilde{\mathcal{B}}\bar{\times}_M\mathcal{X}_t+\mathcal{E}_t).$$

From Behera et al. (2020), if $\rho(\widetilde{\mathcal{A}}_1)<1$ and $\mathcal{Y}_0$ is finite a.s., then $\lim_{t\to\infty}\widetilde{\mathcal{A}}_1^t\bar{\times}_N\mathcal{Y}_0=\mathcal{O}$ and the operator $\sum_{k=0}^{\infty}\widetilde{\mathcal{A}}_1^k L^k$ applied to a sequence $\mathcal{Y}_t$ s.t. $|\mathcal{Y}_{\mathbf{i},t}|<c$ a.s. $\forall\,\mathbf{i}$ converges to the inverse operator $(\mathcal{I}-\widetilde{\mathcal{A}}_1 L)^{-1}$. By the properties of the contracted product we get

$$\begin{aligned}\mathcal{Y}_t&=\sum_{k=0}^{\infty}\widetilde{\mathcal{A}}_1^k\bar{\times}_N(L^k\widetilde{\mathcal{A}}_0)+\sum_{k=0}^{\infty}(\widetilde{\mathcal{A}}_1^k\bar{\times}_N\widetilde{\mathcal{B}})\bar{\times}_M(L^k\mathcal{X}_t)+\sum_{k=0}^{\infty}\widetilde{\mathcal{A}}_1^k\bar{\times}_N(L^k\mathcal{E}_t)\\&=(\mathcal{I}-\widetilde{\mathcal{A}}_1 L)^{-1}\bar{\times}_N\widetilde{\mathcal{A}}_0+\sum_{k=0}^{\infty}\widetilde{\mathcal{A}}_1^k\bar{\times}_N\widetilde{\mathcal{B}}\bar{\times}_M\mathcal{X}_{t-k}+\sum_{k=0}^{\infty}\widetilde{\mathcal{A}}_1^k\bar{\times}_N\mathcal{E}_{t-k}\,.\end{aligned}$$

From the assumption $\mathcal{E}_t\overset{iid}{\sim}\mathcal{N}_{I_1,\ldots,I_N}(\mathcal{O},\Sigma_1,\ldots,\Sigma_N)$, we know that $\mathbb{E}(\mathcal{Y}_t)=\mathcal{Y}_0$, which is finite. Consider the auto-covariance at lag $h\geq 1$. From Theorem S.1.1, we have $\mathbb{E}\big((\mathcal{Y}_t-\mathbb{E}(\mathcal{Y}_t))\circ(\mathcal{Y}_{t-h}-\mathbb{E}(\mathcal{Y}_{t-h}))\big)=\mathbb{E}\big(\mathcal{Y}_t\circ\mathcal{Y}_{t-h}\big)=\mathbb{E}\big(\mathcal{Y}_t\bar{\times}_1\mathcal{Y}_{t-h}^T\big)$. Using the infinite moving average representation for $\mathcal{Y}_t$, we get

$$\begin{aligned}\mathbb{E}\Big(\mathcal{Y}_t\bar{\times}_1\mathcal{Y}_{t-h}^T\Big)&=\mathbb{E}\Big(\big(\sum_{k=0}^{h-1}\mathcal{A}^k\bar{\times}_N\mathcal{E}_{t-k}+\sum_{k=0}^{\infty}\mathcal{A}^{k+h}\bar{\times}_N\mathcal{E}_{t-k-h}\big)\bar{\times}_1\big(\sum_{k=0}^{\infty}\mathcal{A}^k\bar{\times}_N\mathcal{E}_{t-k-h}\big)^T\Big)\\&=\mathbb{E}\Big(\big(\sum_{k=0}^{\infty}\mathcal{A}^{k+h}\bar{\times}_N\mathcal{E}_{t-k-h}\big)\bar{\times}_1\big(\sum_{k=0}^{\infty}\mathcal{E}_{t-k-h}^T\bar{\times}_N(\mathcal{A}^T)^k\big)\Big),\end{aligned}$$

where we used the assumption of independence of $\mathcal{E}_t,\mathcal{E}_{t-h}$, for any $h\geq 0$, and the fact that $(\mathcal{X}\bar{\times}_N\mathcal{Y})^T=(\mathcal{Y}^T\bar{\times}_N\mathcal{X}^T)$. Using $\mathbb{E}(\mathcal{E}_t)=\mathcal{O}$ and linearity of expectation and of the contracted product we get

$$\begin{aligned}\mathbb{E}\Big(\mathcal{Y}_t\bar{\times}_1\mathcal{Y}_{t-h}^T\Big)&=\sum_{k=0}^{\infty}\mathcal{A}^{k+h}\bar{\times}_N\mathbb{E}\Big(\mathcal{E}_{t-k-h}\bar{\times}_1\mathcal{E}_{t-k-h}^T\Big)\bar{\times}_N(\mathcal{A}^T)^k\\&=\sum_{k=0}^{\infty}\mathcal{A}^{k+h}\bar{\times}_N\mathbf{\Sigma}\bar{\times}_N(\mathcal{A}^T)^k=\mathcal{A}^h\bar{\times}_N(\mathcal{I}-\mathcal{A}\bar{\times}_N\mathbf{\Sigma}\bar{\times}_N\mathcal{A}^T)^{-1},\end{aligned}$$

where $\mathbb{E}(\mathcal{E}_{t-k-h}\bar{\times}_1\mathcal{E}_{t-k-h}^T)=\mathbb{E}(\mathcal{E}_{t-k-h}\circ\mathcal{E}_{t-k-h})=\mathbf{\Sigma}=\Sigma_1\circ\ldots\circ\Sigma_N$. From the assumption $\rho(\mathcal{A})<1$ it follows that the above series converges to a finite limit, which is independent from t, thus proving that the process is weakly stationary. $\square$

Proof of Proposition 2.2. From Brazell et al. (2013, Theorem 3.2, Corollary 3.3), we know that $\mathbb{T}$ is a group (called tensor group) and that the matricization operator $\text{mat}_{1:N,1:N}$ is an isomorphism between $\mathbb{T}$ and the linear group of square matrices of size $I^* = \prod_{n=1}^N I_n$. Therefore, there exists a one-to-one relationship between the two eigenvalue problems $\mathcal{A}\bar{\times}_N\mathcal{X} = \lambda\mathcal{X}$ and $A\mathbf{x} = \widetilde{\lambda}\mathbf{x}$, where $A = \text{mat}_{1:N,1:N}(\mathcal{A})$. In particular, $\lambda = \widetilde{\lambda}$ and $\mathbf{x} = \text{vec}\left(\mathcal{X}\right)$. Consequently, $\rho(A) = \rho(\mathcal{A})$ and the result follows for $p = 1$ from the fact that $\rho(A) < 1$ is a sufficient condition for the VAR(1) stationarity Lütkepohl (2005, Proposition 2.1). Since any VAR(p) and ART(p) processes can be rewritten as VAR(1) and ART(1), respectively, on an augmented state space, the result follows for any $p \geq 1$. □

Proof of Lemma 2.1. Consider a ART(p) process with $\mathcal{Y}_t \in \mathbb{R}^{I_1\times\ldots\times I_N}$ and $p \geq 1$. We define the $(pI_1 \times I_2 \times \ldots \times I_N)$-dimensional tensors $\underline{\mathcal{Y}}_t$ and $\underline{\mathcal{E}}_t$ as $\underline{\mathcal{Y}}_{(k-1)I_1+1:kI_1,:,\ldots,:,t} = \mathcal{Y}_{t-k}$ and $\underline{\mathcal{E}}_{(k-1)I_1+1:kI_1,:,\ldots,:,t} = \mathcal{E}_{t-k}$, for $k = 0,\ldots,p$, respectively. Define the $(pI_1 \times I_2 \times \ldots \times I_N \times pI_1 \times I_2 \times \ldots \times I_N)$-dimensional tensor $\underline{\mathcal{A}}$ as $\underline{\mathcal{A}}_{(1:I_1,:,\ldots,:,(k-1)I_1+1:kI_1,:,\ldots,:} = \mathcal{A}_k$, for $k = 1,\ldots,p$, $\underline{\mathcal{A}}_{(kI_1+1:(k+1)I_1,:,\ldots,:,(k-1)I_1+1:kI_1,:,\ldots,:} = \mathcal{I}$, for $k = 1,\ldots,p-1$ and 0 elsewhere. Using this notation, we can rewrite the $(I_1 \times I_2 \times \ldots \times I_N)$-dimensional ART($p$) process $\mathcal{Y}_t = \sum_{k=1}^p \mathcal{A}_k\bar{\times}_N\mathcal{Y}_{t-j} + \mathcal{E}_t$ as the $(pI_1 \times I_2 \times \ldots \times I_N)$-dimensional ART(1) process $\underline{\mathcal{Y}}_t = \underline{\mathcal{A}}\bar{\times}_N\underline{\mathcal{Y}}_{t-1} + \underline{\mathcal{E}}_t$. □

S.4.1 Special cases and corresponding proofs

The model in eq. (5) is a generalization of several well-known econometric models, as shown in the following remarks.

Remark S.4.1 (Univariate)**.** *If $I_i = 1$ for $i = 1,\ldots,N$, then model (5) reduces to a univariate regression*

$$y_t = \alpha_0 + \sum_{j=1}^p \alpha_j y_{t-j} + \boldsymbol{\beta}' \text{vec}\left(\mathcal{X}_t\right) + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0,\sigma^2), \tag{S10}$$

where the coefficients of (5) become $\mathcal{A}_j = \alpha_j \in \mathbb{R}$, $j = 0,\ldots,p$ and $\mathcal{B} = \boldsymbol{\beta} \in \mathbb{R}^{J^*}$.

Proof. Consider model (5) when $I_j = 1$, for $j = 1,\ldots,N$. Note that a N-order tensor whose modes have all unit length is equivalent to a 1-order tensor, i.e. a scalar. As a consequence, the dependent variable becomes $y_t \in \mathbb{R}$ and the autoregressive coefficient tensors reduce to $\alpha_j \in \mathbb{R}$, $j = 0, \ldots, p$. The coefficient tensor related to the covariates $\mathcal{X}_t$ becomes a vector $\boldsymbol{\beta} \in \mathbb{R}^{J^*}$. Finally, the error term distribution reduces to a univariate normal with 0 mean and variance σ^2. In this framework, the mode-$N+1$ product reduces to the standard inner product between vectors.

The PARAFAC(R) decomposition can still be applied in this case. We get

$$\alpha_j = \sum_{r=1}^{R} \alpha_{j,1}^{(r)} \circ \ldots \circ \alpha_{j,N}^{(r)} = \sum_{r=1}^{R} \alpha_{j,1}^{(r)} \cdots \alpha_{j,N}^{(r)},$$

for each $j = 0, \ldots, p$, where the outer product reduces to the ordinary scalar multiplication and all $\alpha_{j,k}^{(r)}$, $k = 1, \ldots, N$, $r = 1, \ldots, R$ are scalars. Similarly, we have

$$\boldsymbol{\beta} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \ldots \circ \beta_N^{(r)} \circ \boldsymbol{\beta}_{N+1}^{(r)} = \sum_{r=1}^{R} \beta_1^{(r)} \cdot \ldots \cdot \beta_N^{(r)} \cdot \boldsymbol{\beta}_{N+1}^{(r)}$$

since again the outer product reduces to the ordinary scalar multiplication and all $\beta_{j,k}^{(r)}$, $k = 1, \ldots, N$, $r = 1, \ldots, R$ are scalars, while the marginal corresponding to the last mode $N+1$ is a vector of length J^*. □

Remark S.4.2 (SUR)**.** *If $I_i = 1$ for $i = 2, \ldots, N$ and define by $\mathbf{1}_n$ the unit vector of length n, then model (5) reduces to a Seemingly Unrelated Regression (SUR) model (Zellner, 1962)*

$$\mathbf{y}_t = \boldsymbol{\alpha}_0 + B \times_2 \operatorname{vec}\left(\mathcal{X}_t\right) + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m(\mathbf{0}, \Sigma), \tag{S11}$$

where $I_1 = m$ and the coefficients of (5) become $\mathcal{A}_j = 0$, $j = 1, \ldots, p$, $\mathcal{A}_0 = \boldsymbol{\alpha}_0 \in \mathbb{R}^m$ and $\mathcal{B} = B \in \mathbb{R}^{m \times J^*}$. Note that, by definition, $B \times_2 \operatorname{vec}\left(\mathcal{X}_t\right) = B \operatorname{vec}\left(\mathcal{X}_t\right)$.

Remark S.4.3 (VARX and Panel VAR)**.** *Consider the setup of Remark S.4.2. If $\mathbf{z}_t = \mathbf{y}_{t-1}$, then we obtain a VARX(1) model, with restricted covariance matrix. Another vector of regressors $\mathbf{w}_t = \operatorname{vec}\left(W_t\right) \in \mathbb{R}^q$ may enter the regression (S11) pre-multiplied (along mode-3) by a tensor $\mathcal{D} \in \mathbb{R}^{m \times n \times q}$. Therefore, model (5) encompasses as a particular case also the panel VAR models of Canova and Ciccarelli (2004, 2009); Canova et al. (2007), provided that we make the same restriction on Σ.*

Proof. Consider model (5)s with $I_1 = m$ and $I_j = 1$, for $j = 2, \ldots, N$. Denote by $\mathbf{x}_t = \operatorname{vec}\left(\mathcal{X}_t\right)$ the external covariates. Note that the mode-$N+1$ product become mode-2 product and the distribution of the error term reduces to the multivariate (m-dimensional) normal. The dependent variable reduces to the vector $\mathbf{y}_t \in \mathbb{R}^m$ while the coefficient tensors become $\boldsymbol{\alpha}_0 \in \mathbb{R}^m$, $A_j \in \mathbb{R}^{m \times m}$, for $j = 1, \ldots, p$ and $B \in \mathbb{R}^{m \times J^*}$.

Assuming a PARAFAC(R) decomposition, we get the same result for $\boldsymbol{\alpha}_0$ as in the previous proof, having in this case $N-1$ scalar marginals and one vector marginal. For the remaining tensors, it holds

$$A_j = \sum_{r=1}^{R} \boldsymbol{\alpha}_{j,1}^{(r)} \circ \left(\alpha_{j,2}^{(r)} \cdot \ldots \cdot \alpha_{j,N-1}^{(r)}\right) \circ \boldsymbol{\alpha}_N^{(r)} = \sum_{r=1}^{R} A_j^{(r)} \cdot \left(\alpha_{j,2}^{(r)} \cdot \ldots \cdot \alpha_{j,N-1}^{(r)}\right).$$

Similarly, for the matrix B one gets

$$B = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \left(\beta_2^{(r)} \cdot \ldots \cdot \beta_N^{(r)}\right) \circ \boldsymbol{\beta}_{N+1}^{(r)} = \sum_{r=1}^{R} B^{(r)} \cdot \left(\beta_2^{(r)} \cdot \ldots \cdot \beta_N^{(r)}\right).$$

It remains to prove that the structure imposed by standard VARX and Panel VAR models holds also in the model of eq. (5). Notice that the latter does not impose any restriction on the coefficients, other than the PARAFAC(R) decomposition. It must be stressed that it is not possible to achieve the desired structure of the coefficients, in terms of the location of the zeros, by means of an accurate choice of the marginals. In fact, the decomposition we are assuming does not allow to create a particular structure on the resulting tensor.

Nonetheless, it is still possible to achieve the desired result by a slight modification of the model in eq. (5). For example, consider the coefficient tensor $\mathcal{B}$, then to create a tensor whose entries are non-zero only in some pre-specified (hence *a-priori* known) cells, it suffices to multiply $\mathcal{B}$ by a binary tensor (i.e. one where all entries are either 0 or 1) via the Hadamard product. In formulas, let $\mathcal{H} \in \{0,1\}^{I_1 \times \ldots \times I_N \times J}$, such that it has 0 only in those cells which are known to be null. Then $\bar{\mathcal{B}} = \mathcal{H} \odot \mathcal{B}$ has the desired structure. The same way of reasoning holds for any coefficient tensor as well as for the covariance matrices.

To conclude, in Panel VAR models one generally has as regressors in each equation a function of the endogenous variables (for example their average). Since this does not affect the coefficients of the model, it is possible to re-create it in our framework by simply rearranging the regressors in eq. (5) accordingly. In terms of the model, none of the issues described invalidates the formulation of eq. (5), which is able to encompass all of them by suitable rearrangements of the covariates and/or the coefficients, which are consistent with the general model. $\square$

Remark S.4.4 (VECM)**.** *The model in eq. (5) generalises the Vector Error Correction Model (VECM) widely used in multivariate time series analysis (Engle and Granger, 1987; Schotman and Van Dijk, 1991). Consider a K-dimensional VAR(1) model*

$$\mathbf{y}_t = B\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m(\mathbf{0}, \Sigma).$$

Defining $\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ and $\Pi = (B - I) = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $K \times R$ matrices of rank $R < K$, we obtain the associated VECM

$$\Delta\mathbf{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t. \tag{S12}$$

This is used for studying the cointegration relations among the components of $\mathbf{y}_t$. Since $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}' = \sum_{r=1}^R \boldsymbol{\alpha}_{:,r}\boldsymbol{\beta}'_{:,r} = \sum_{r=1}^R \tilde{\boldsymbol{\beta}}_1^{(r)} \circ \tilde{\boldsymbol{\beta}}_2^{(r)}$, we can interpret the VECM model in eq. (S12) as a particular case of the model in eq. (5) where the coefficient $\mathcal{B}$ is the matrix $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}'$. Furthermore by writing $\Pi = \sum_{r=1}^R \tilde{\boldsymbol{\beta}}_1^{(r)} \circ \tilde{\boldsymbol{\beta}}_2^{(r)}$ we can interpret this relation as a rank-R PARAFAC decomposition of $\mathcal{B}$. Following this analogy, the PARAFAC rank corresponds to the cointegration rank, $\tilde{\boldsymbol{\beta}}_1^{(r)}$ are the mean-reverting coefficients and $\tilde{\boldsymbol{\beta}}_2^{(r)} = (\tilde{\beta}_{2,1}^{(r)}, \ldots, \tilde{\beta}_{2,K}^{(r)})$ are the cointegrating vectors. See Section S.4 for details. This interpretation opens the way to reparametrization of $\mathcal{B}$ based on tensor SVD representations, and to the application of regularization methods in the spirit of Baştürk et al. (2017). This is beyond the scope of the paper, thus we leave it for further research.

Remark S.4.5 (follows from Remark S.4.4)**.** *From the VECM in eq. (S12) and denoting $\mathbf{y}_{t-1} = \mathrm{vec}\left(Y_{t-1}\right)$ we can obtain an explicit form for the long run equilibrium (or cointegrating) relations, as follows*

$$\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} = \Big(\sum_{r=1}^R \boldsymbol{\gamma}_1^{(r)} \circ \boldsymbol{\gamma}_2^{(r)}\Big) \bar{\times}_1 \mathbf{y}_{t-1} = \Big(\sum_{r=1}^R \boldsymbol{\gamma}_1^{(r)}\boldsymbol{\gamma}_2^{(r)\prime}\Big)\mathbf{y}_{t-1} = \sum_{r=1}^R \boldsymbol{\gamma}_1^{(r)}(\boldsymbol{\gamma}_2^{(r)\prime}\mathbf{y}_{t-1}),$$

where $\boldsymbol{\gamma}_1^{(r)}$ and $\boldsymbol{\gamma}_2^{(r)}$ are vectors of length K. The marginals $(\boldsymbol{\gamma}_2^{(r)})_r$ can thus be interpreted as thelong run cointegrating relationships, and the marginals $(\boldsymbol{\gamma}_2^{(r)})_r$ are the corresponding loadings.

Remark S.4.6 (MAI of Carriero et al. (2016)). *The multivariate autoregressive index model (MAI) of Carriero et al. (2016) is another special case of model (5). A MAI is a VAR model with a low rank decomposition imposed on the coefficient matrix, as follows*

$$\mathbf{y}_t = \mathbf{A}\mathbf{B}_0\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\mathbf{y}_t$ is a $(n\times 1)$ vector, whereas $\mathbf{A},\mathbf{B}_0$ are $(n\times R)$ and $(R\times n)$ matrices, respectively. In Carriero et al. (2016), the authors assumed $R=1$. This corresponds to our parametrization using $R=1$ and defining $\mathbf{A}\boldsymbol{\beta}_1^{(1)}$ and $\mathbf{B}_0' = \boldsymbol{\beta}_2^{(1)}$, which leads us to $\mathbf{A}\mathbf{B}_0 = \boldsymbol{\beta}_1^{(1)} \circ \boldsymbol{\beta}_2^{(1)}$.

S.5 Prior distribution on tensor entries

Proof of Lemma 3.1. The distribution of each of these products has been characterised by Springer and Thompson (1970), who proved the following theorem.

Theorem S.5.1 (4 in Springer and Thompson (1970)). *The probability density function of the product $z = \prod_{h=1}^H x_h$ of H independent Normal random variables $x_h \sim \mathcal{N}(0,\sigma_h^2)$, $h=1,\dots,H$, is proportional to a Meijer G-function*

$$p(z|(\sigma_h^2)_{h=1}^H) = K\cdot G_{H,0}^{H,0}\left(z^2 \prod_{h=1}^H \frac{1}{2\sigma_h} \middle| \mathbf{0} \right),$$

where the normalising constant is

$$K = \left((2\pi)^{H/2} \prod_{h=1}^H \sigma_h \right)^{-1}$$

and $G_{p,q}^{m,n}(\cdot|\cdot)$ is a Meijer G-function (with $c\in\mathbb{R}$ and $s\in\mathbb{C}$)

$$G_{p,q}^{m,n}\left(z \middle| \begin{matrix} a_1,\dots,a_p \\ b_1,\dots,b_q \end{matrix} \right) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^{-s} \frac{\prod_{j=1}^m \Gamma(s+b_j)\cdot \prod_{j=1}^n \Gamma(1-a_j-s)}{\prod_{j=n+1}^p \Gamma(s+a_j)\cdot \prod_{j=m+1}^q \Gamma(1-b_j-s)}\,\mathrm{d}s.$$

The integral is taken over a vertical line in the complex plane. Note that in the special case $H=2$ we have $z \sim c_1P_1 - c_2P_2$, with $P_1,P_2 \sim \chi_1^2$ and $c_1 = Var(x_1+x_2)/4$, $c_2 = Var(x_1-x_2)/4$. In this case, the resulting distribution is called product Normal distribution.

Therefore, the result follows from Theorem S.5.1, with $z = \beta_r$, $H = 4$, $\sigma_h = \tau\phi_r w_{h,r,m_h}$ and where the parameters of the G-function are $m = p = 4 =$, $n = q = 0$, $(a_1, \dots, a_p) = (0, \dots, 0)$ and $(b_1, \dots, b_q) = (0, \dots, 0)$. $\square$

We assessed the shape of this marginal distribution in a simulated setting, and found that it has fatter tails than the Gaussian distribution. In particular, Fig. 4 show the empirical distribution of two randomly chosen entries of a 3-order tensor $\mathcal{B}$ whose PARAFAC decomposition is assumed with $R = 5$. The probability density function of a Laplace (or double exponential) distribution with mean $\mu \in \mathbb{R}$ and variance $2b^2$, with $b > 0$, is

$$f(x|\mu, b) = \frac{1}{2b} \exp\Big(- \frac{|x - \mu|}{2b} \Big) \qquad x \in \mathbb{R}.$$

Compared to the standard normal and standard Laplace distribution, the prior distribution induced on the single entries of the tensor has fatter tails.

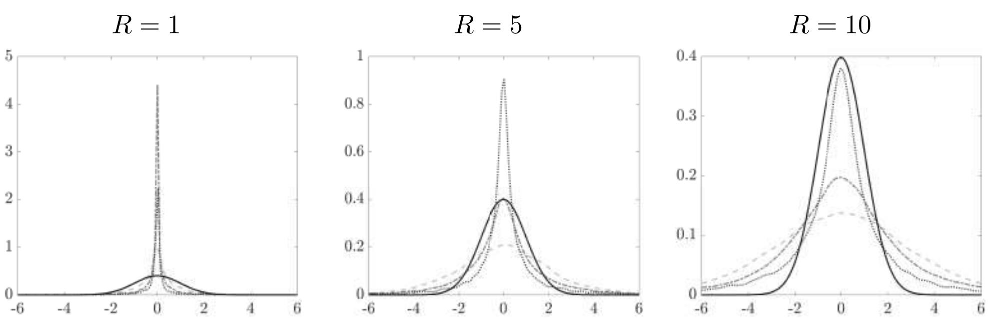

Figure 2: Monte Carlo simulation from the prior distribution of entry $b_{i_1,\dots,i_N}$ of a generic N-order tensor, for varying rank R. In column: simulation with $R = 1$ (left), $R = 5$ (middle) and $R = 10$ (right). In all plots: standard Normal (continuous line) and prior for $b_{i_1,\dots,i_N}$, for $N = 2$ (dashed line), $N = 4$ (dash-dotted line) and $N = 6$ (dotted line).

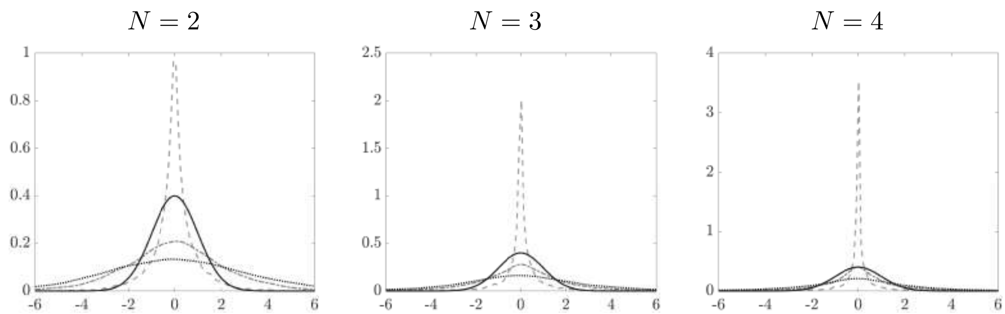

Figure 3: Monte Carlo simulation from the prior distribution of entry $b_{i_1,\ldots,i_N}$ of a N-order tensor, with rank R, for varying N. In column: simulation with $N = 2$ (left), $N = 3$ (middle) and $N = 4$ (right). In all plots: standard Normal (continuous line) and prior for $b_{i_1,\ldots,i_N}$, for $R = 1$ (dashed line), $R = 5$ (dash-dotted line) and $R = 10$ (dotted line).

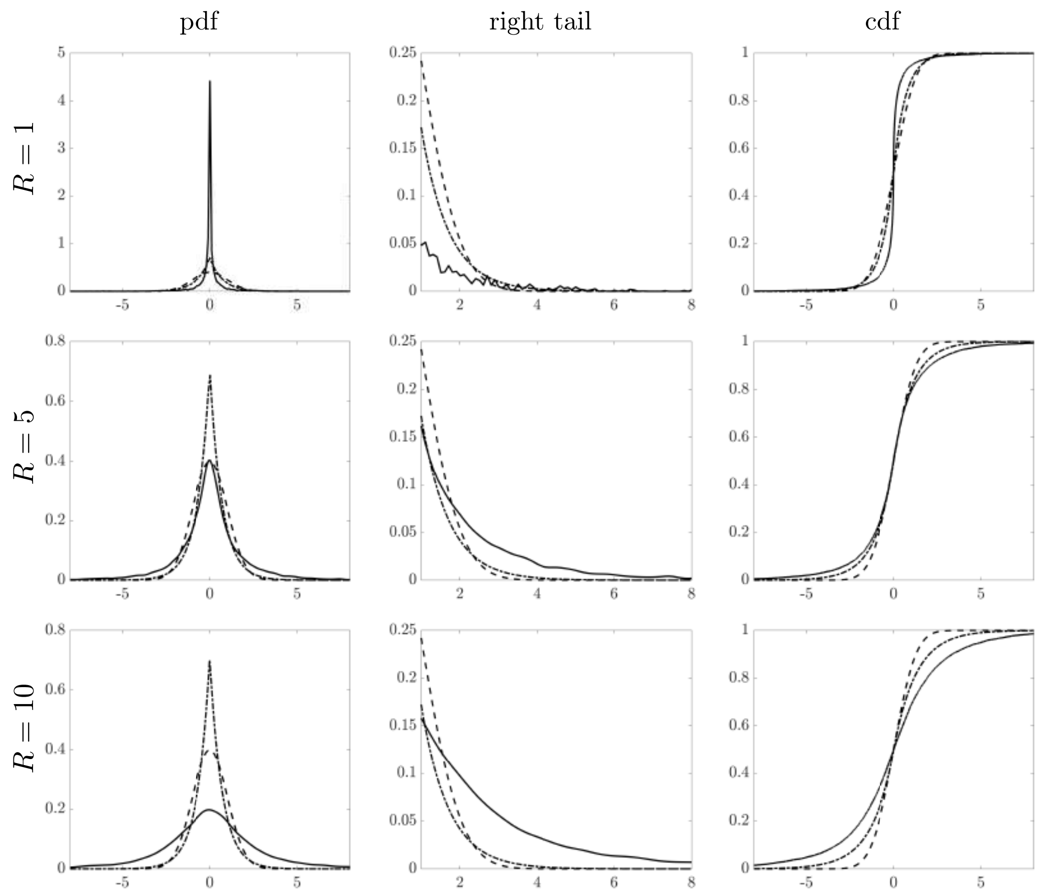

Figure 4: Monte Carlo simulation from the prior distribution of a generic 4-order tensor entry b_{ijkp} (continuous line), standard Normal distribution (dashed line) and standard Laplace distribution (dash-dotted line). In column: probability density function (*left*), right tail the probability density function (*middle*), cumulative distribution function (*right*). In row: simulations with $R = 1$, $R = 5$ and $R = 10$ (*first*, *second* and *third*, respectively).

S.6 Computational details - ART(1) model

In this section we will follow the convention of denoting the prior distributions with $\pi(\cdot)$. In addition, let $\mathbf{W} = (W_{j,r})_{j,r}$ be the collection of all (local variance) matrices $W_{j,r}$, for $j = 1, \ldots, J$ and $r = 1, \ldots, R$, let $I_0 = \sum_{j=1}^{J} I_j$ be the sum of the length of each mode of the tensor $\mathcal{B}$ and let $\mathbf{Y} = (\mathcal{Y}_t)_t$ the collection of observed variables. Recall that in the ART(1) model in eq. (15), the variable $\mathcal{Y}_t$ is a 3-order tensor, thus we have $J = 4$.

S.6.1 Full conditional distribution of ϕ_r

In order to derive this posterior distribution, we use Guhaniyogi et al. (2017, Lemma 7.9). Recall that: $a_\tau = \alpha R$ and $b_\tau = \alpha(R)^{1/J}$. The posterior full conditional distribution of $\boldsymbol{\phi}$ is

$$
\begin{aligned}
p(\boldsymbol{\phi}|\mathcal{B},\mathbf{W}) &\propto \pi(\boldsymbol{\phi}) \int_0^{+\infty} p(\mathcal{B}|\mathbf{W},\boldsymbol{\phi},\tau)\pi(\tau)\mathrm{d}\tau \\
&\propto \prod_{r=1}^{R} \phi_r^{\alpha-1} \int_0^{+\infty} \Big(\prod_{r=1}^{R}\prod_{j=1}^{J} (\tau\phi_r)^{-I_j/2} \big|W_{j,r}\big|^{-1/2} \\
&\quad \cdot \exp\Big(- \frac{1}{2\tau\phi_r} \boldsymbol{\beta}_j^{(r)\prime} W_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)} \Big)\Big) \cdot \tau^{a_\tau - 1} e^{-b_\tau \tau} \mathrm{d}\tau \\
&\propto \prod_{r=1}^{R} \phi_r^{\alpha-1} \int_0^{+\infty} \Big(\prod_{r=1}^{R} (\tau\phi_r)^{-I_0/2} \exp\Big(- \frac{1}{2\tau\phi_r} \sum_{j=1}^{J} \boldsymbol{\beta}_j^{(r)\prime} W_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)} \Big)\Big) \\
&\quad \cdot \tau^{a_\tau - 1} e^{-b_\tau \tau} \mathrm{d}\tau.
\end{aligned}
$$

Define $C_r = \sum_{j=1}^{J} \boldsymbol{\beta}_j^{(r)\prime} W_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)}$, then group together the powers of τ and ϕ_r as follows

$$
\begin{aligned}
p(\boldsymbol{\phi}|\mathcal{B},\mathbf{W}) &\propto \prod_{r=1}^{R} \phi_r^{\alpha-1-\frac{I_0}{2}} \int_0^{+\infty} \tau^{a_\tau - 1 - \frac{RI_0}{2}} e^{-b_\tau \tau} \Big(\prod_{r=1}^{R} \exp\Big(- \frac{1}{2\tau\phi_r} C_r \Big)\Big) \mathrm{d}\tau \\
&= \prod_{r=1}^{R} \phi_r^{\alpha-1-\frac{I_0}{2}} \int_0^{+\infty} \tau^{a_\tau - 1 - \frac{Rd_0}{2}} \exp\Big(- b_\tau \tau - \sum_{r=1}^{R} \frac{C_r}{2\tau\phi_r} \Big) \mathrm{d}\tau. \qquad \text{(S13)}
\end{aligned}
$$

The probability density function of a Generalized Inverse Gaussian in the parametrization with three parameters ($a > 0$, $b > 0$, $c \in \mathbb{R}$), with $x \in (0, +\infty)$, is given by

$$
x \sim \mathrm{GiG}(a,b,c) \iff p(x|a,b,c) = \frac{(a/b)^{\frac{c}{2}}}{2K_c(\sqrt{ab})} x^{c-1} \exp\Big(- \frac{1}{2}(ax + b/x) \Big),
$$

with $K_c(\cdot)$ a modified Bessel function of the second type. Our goal is to reconcile eq. (S13) to the kernel of this distribution. Since by definition $\sum_{r=1}^{R} \phi_r = 1$, it holds that $\sum_{r=1}^{R} (b_\tau \tau \phi_r) = (b_\tau \tau) \sum_{r=1}^{R} \phi_r = b_\tau \tau$. This allows to rewrite the exponential as

$$
\begin{aligned}
p(\boldsymbol{\phi}|\mathcal{B},\mathbf{W}) &\propto \prod_{r=1}^{R} \phi_r^{\alpha-1-\frac{I_0}{2}} \int_0^{+\infty} \tau^{\left(a_\tau - \frac{RI_0}{2}\right)-1} \exp\Big(- \sum_{r=1}^{R} \Big(\frac{C_r}{2\tau\phi_r} + b_\tau \tau \phi_r \Big)\Big) \mathrm{d}\tau \\
&= \int_0^{+\infty} \Big(\prod_{r=1}^{R} \phi_r^{\alpha - \frac{I_0}{2} - 1} \Big) \tau^{\left(\alpha R - \frac{RI_0}{2}\right)-1} \exp\Big(- \sum_{r=1}^{R} \Big(\frac{C_r}{2\tau\phi_r} + b_\tau \tau \phi_r \Big)\Big) \mathrm{d}\tau
\end{aligned}
$$

where we expressed $a_\tau = \alpha R$. According to the results in Appendix A and Guhaniyogi et al. (2017), the function in the previous equation is the kernel of a generalized inverse

Gaussian for $\psi_r = \tau\phi_r$, which yields the distribution of ϕ_r after normalization. Hence, for $r = 1, \dots, R$, we sample

$$p(\psi_r|\mathcal{B}, \mathbf{W}, \tau, \alpha) \sim \mathrm{GiG}\Big(\alpha - \frac{I_0}{2}, 2b_\tau, 2C_r\Big)$$

then, we obtain ϕ_r by renormalizing (see Kruijer et al., 2010): $\phi_r = \psi_r / \sum_{l=1}^{R} \psi_l$.

S.6.2 Full conditional distribution of τ

The posterior distribution of the global variance parameter, τ, is derived by simple application of Bayes' Theorem

$$\begin{aligned} p(\tau|\mathcal{B}, \mathbf{W}, \boldsymbol{\phi}) &\propto \pi(\tau) p(\mathcal{B}|\mathbf{W}, \boldsymbol{\phi}, \tau) \\ &\propto \tau^{a_\tau - 1} e^{-b_\tau \tau} \Big(\prod_{r=1}^{R} (\tau\phi_r)^{-\frac{I_0}{2}} \exp\Big(-\frac{1}{2\tau\phi_r} \sum_{j=1}^{4} \boldsymbol{\beta}_j^{(r)\prime} (W_{j,r})^{-1} \boldsymbol{\beta}_j^{(r)} \Big) \Big) \\ &\propto \tau^{a_\tau - \frac{RI_0}{2} - 1} \exp\Big(-b_\tau \tau - \Big(\sum_{r=1}^{R} \frac{C_r}{\phi_r} \frac{1}{\tau} \Big) \Big). \end{aligned}$$

This is the kernel of a generalized inverse Gaussian

$$p(\tau|\mathcal{B}, \mathbf{W}, \boldsymbol{\phi}) \sim \mathrm{GiG}\Big(a_\tau - \frac{RI_0}{2}, 2b_\tau, 2\sum_{r=1}^{R} \frac{C_r}{\phi_r}\Big).$$

S.6.3 Full conditional distribution of $\lambda_{j,r}$

Start by observing that, for $j = 1, \dots, 4$ and $r = 1, \dots, R$, the prior distribution on the vector $\boldsymbol{\beta}_j^{(r)}$ defined in eq. (17) implies that each component follows a double exponential distribution

$$\beta_{j,p}^{(r)} \sim DE\Big(0, \frac{\lambda_{j,r}}{\sqrt{\tau\phi_r}}\Big)$$

with probability density function given by

$$\pi(\beta_{j,p}^{(r)}|\lambda_{j,r}, \phi_r, \tau) = \frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}} \exp\Big(-\frac{|\beta_{j,p}^{(r)}|}{(\lambda_{j,r}/\sqrt{\tau\phi_r})^{-1}} \Big). \tag{S14}$$

Then, exploiting the Gamma prior and eq. (S14)

$$p(\lambda_{j,r}|\boldsymbol{\beta}_j^{(r)}, \phi_r, \tau) \propto \pi(\lambda_{j,r}) p(\boldsymbol{\beta}_j^{(r)}|\lambda_{j,r}, \phi_r, \tau)$$

$$
\begin{aligned}
&\propto \lambda_{j,r}^{a_\lambda - 1} e^{-b_\lambda \lambda_{j,r}} \prod_{p=1}^{I_j} \frac{\lambda_{j,r}}{2\sqrt{\tau \phi_r}} \exp\Big(- \frac{\big|\beta_{j,p}^{(r)}\big|}{(\lambda_{j,r}/\sqrt{\tau\phi_r})^{-1}} \Big) \\
&= \lambda_{j,r}^{a_\lambda - 1} \Big(\frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}} \Big)^{I_j} e^{-b_\lambda \lambda_{j,r}} \exp\Big(- \frac{\sum_{p=1}^{I_j} \big|\beta_{j,p}^{(r)}\big|}{\sqrt{\tau\phi_r}/\lambda_{j,r}} \Big) \\
&\propto \lambda_{j,r}^{(a_\lambda + I_j) - 1} \exp\Bigg(- \Bigg(b_\lambda + \frac{\big\|\boldsymbol{\beta}_j^{(r)}\big\|_1}{\sqrt{\tau\phi_r}} \Bigg) \lambda_{j,r} \Bigg).
\end{aligned}
$$

Thus, the full conditional distribution of $\lambda_{j,r}$ is given by

$$
p(\lambda_{j,r}|\mathcal{B}, \phi_r, \tau) \sim \mathcal{G}a\Bigg(a_\lambda + I_j, b_\lambda + \frac{\big\|\boldsymbol{\beta}_j^{(r)}\big\|_1}{\sqrt{\tau\phi_r}} \Bigg).
$$

S.6.4 Full conditional distribution of $w_{j,r,p}$

We sample independently each component $w_{j,r,p}$ of the matrix $W_{j,r} = \operatorname{diag}(\mathbf{w}_{j,r})$, for $p = 1, \ldots, I_j$, $j = 1, \ldots, 4$ and $r = 1, \ldots, R$, from the full conditional distribution

$$
\begin{aligned}
p(w_{j,r,p}|\boldsymbol{\beta}_j^{(r)}, &\lambda_{j,r}, \phi_r, \tau) \propto p(\beta_{j,p}^{(r)}|w_{j,r,p}, \phi_r, \tau)\pi(w_{j,r,p}|\lambda_{j,r}) \\
&= (\tau\phi_r)^{-\frac{1}{2}} w_{j,r,p}^{-\frac{1}{2}} \exp\Big(- \frac{1}{2\tau\phi_r} \beta_{j,p}^{(r)^2} w_{j,r,p}^{-1} \Big) \frac{\lambda_{j,r}^2}{2} \exp\Big(- \frac{\lambda_{j,r}^2}{2} w_{j,r,p} \Big) \\
&\propto w_{j,r,p}^{-\frac{1}{2}} \exp\Big(- \frac{\lambda_{j,r}^2}{2} w_{j,r,p} - \frac{\beta_{j,p}^{(r)^2}}{2\tau\phi_r} w_{j,r,p}^{-1} \Big),
\end{aligned}
$$

where the second row comes from the fact that $w_{j,r,p}$ influences only the p-th component of the vector $\boldsymbol{\beta}_j^{(r)}$. Hence, we get

$$
p(w_{j,r,p}|\boldsymbol{\beta}_j^{(r)}, \lambda_{j,r}, \phi_r, \tau) \sim \operatorname{GiG}\Big(\frac{1}{2}, \lambda_{j,r}^2, \frac{\beta_{j,p}^{(r)^2}}{\tau\phi_r} \Big).
$$

S.6.5 Full conditional distributions of PARAFAC marginals

Define $\boldsymbol{\alpha}_1 \in \mathbb{R}^I$, $\boldsymbol{\alpha}_2 \in \mathbb{R}^J$ and $\boldsymbol{\alpha}_3 \in \mathbb{R}^K$ and let $\mathcal{A} = \operatorname{vec}\big(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2 \circ \boldsymbol{\alpha}_3\big)$. Then, from Theorem S.1.3 it holds

$$
\begin{aligned}
\operatorname{vec}\big(\mathcal{A}\big) &= \operatorname{vec}\big(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2 \circ \boldsymbol{\alpha}_3\big) = \boldsymbol{\alpha}_3 \otimes \operatorname{vec}\big(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\big) \\
&= \boldsymbol{\alpha}_3 \otimes \big(\boldsymbol{\alpha}_2 \otimes \mathbf{I}_I\big) \operatorname{vec}\big(\boldsymbol{\alpha}_1\big) = \big(\boldsymbol{\alpha}_3 \otimes \boldsymbol{\alpha}_2 \otimes \mathbf{I}_I\big)\boldsymbol{\alpha}_1 \qquad \text{(S15)} \\
&= \boldsymbol{\alpha}_3 \otimes \big(\big(\mathbf{I}_J \otimes \boldsymbol{\alpha}_1\big) \operatorname{vec}\big(\boldsymbol{\alpha}_2'\big)\big) = \big(\boldsymbol{\alpha}_3 \otimes \mathbf{I}_J \otimes \boldsymbol{\alpha}_1\big)\boldsymbol{\alpha}_2 \qquad \text{(S16)}
\end{aligned}
$$

$$
\begin{aligned}
&= \operatorname{vec}\big(\operatorname{vec}\big(\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2'\big)\boldsymbol{\alpha}_3'\big) = \big(\mathbf{I}_K \otimes \operatorname{vec}\big(\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2'\big)\big)\operatorname{vec}\big(\boldsymbol{\alpha}_3'\big) \\
&= \big(\mathbf{I}_K \otimes \operatorname{vec}\big(\boldsymbol{\alpha}_1\boldsymbol{\alpha}_2'\big)\big)\boldsymbol{\alpha}_3 = \big(\mathbf{I}_K \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_1\big)\boldsymbol{\alpha}_3. \qquad \text{(S17)}
\end{aligned}
$$

Consider the model in eq. (15), it holds

$$
\begin{aligned}
\mathcal{Y}_t &= \mathcal{B}\bar{\times}_1 \mathbf{x}_t + \mathcal{E}_t \\
\operatorname{vec}\big(\mathcal{Y}_t\big) &= \operatorname{vec}\big(\mathcal{B}\bar{\times}_1 \mathbf{x}_t + \mathcal{E}_t\big) \\
&= \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1 \mathbf{x}_t\big) + \operatorname{vec}\big(\mathcal{B}_r\bar{\times}_1 \mathbf{x}_t\big) + \operatorname{vec}\big(\mathcal{E}_t\big),
\end{aligned}
$$

where the term in the middle can be re-written as

$$
\operatorname{vec}\big(\mathcal{B}_r\bar{\times}_1 \mathbf{x}_t\big) = \operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\big) \cdot \mathbf{x}_t'\boldsymbol{\beta}_4^{(r)}.
$$

It is then possible to make explicit the dependence on each PARAFAC marginal by exploiting the results in eq. (S15)-(S17), as follows

$$
\begin{aligned}
\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\big) \cdot \mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} &= \operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\big) \cdot \mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} = \mathbf{b}_4\boldsymbol{\beta}_4^{(r)} && \text{(S18)} \\
&= \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \mathbf{I}_I\big)\boldsymbol{\beta}_1^{(r)} = \mathbf{b}_1\boldsymbol{\beta}_1^{(r)} && \text{(S19)} \\
&= \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)} \otimes \mathbf{I}_J \otimes \boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_2^{(r)} = \mathbf{b}_2\boldsymbol{\beta}_2^{(r)} && \text{(S20)} \\
&= \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t\rangle\big(\mathbf{I}_K \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_3^{(r)} = \mathbf{b}_3\boldsymbol{\beta}_3^{(r)}. && \text{(S21)}
\end{aligned}
$$

Given a sample of length T and assuming that the distribution at time $t = 0$ is known (as standard practice in time series analysis), the likelihood function is

$$
\begin{aligned}
L(\mathbf{Y}|\mathcal{B}, \Sigma_1, \Sigma_2, \Sigma_3) &= \prod_{t=1}^{T} (2\pi)^{-\frac{I_1 I_2 I_3}{2}} \big|\Sigma_3\big|^{-\frac{I_1 I_2}{2}} \big|\Sigma_2\big|^{-\frac{I_1 I_3}{2}} \big|\Sigma_1\big|^{-\frac{I_2 I_3}{2}} \\
&\quad \cdot \exp\Big(-\frac{1}{2}(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t)\bar{\times}_3\big(\circ_{j=1}^{3} \Sigma_j^{-1}\big)\bar{\times}_3(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\sum_{t=1}^{T} \tilde{\mathcal{E}}_t\bar{\times}_3(\Sigma_1^{-1} \circ \Sigma_2^{-1} \circ \Sigma_3^{-1})\bar{\times}_3\tilde{\mathcal{E}}_t\Big),
\end{aligned}
$$

with

$$
\begin{aligned}
\operatorname{vec}\big(\tilde{\mathcal{E}}_t\big) &= \operatorname{vec}\big(\mathcal{Y}_t - \mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t - \big(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\big)\langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t\rangle\big) \\
&= \operatorname{vec}\big(\mathcal{Y}_t\big) - \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big) - \operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\big)\langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t\rangle.
\end{aligned}
$$

Alternatively, by exploiting the relation between the tensor normal distribution and the multivariate normal distribution, we have

$$
\begin{aligned}
L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3) &= \prod_{t=1}^{T}(2\pi)^{-\frac{I_1 I_2 I_3}{2}}\left|\Sigma_3\otimes\Sigma_2\otimes\Sigma_1\right|^{-\frac{1}{2}} \\
&\quad \cdot \exp\Big(-\frac{1}{2}\operatorname{vec}\big(\mathcal{Y}_t-\mathcal{B}\bar{\times}_1\mathbf{x}_t\big)'\big(\Sigma_3^{-1}\otimes\Sigma_2^{-1}\otimes\Sigma_1^{-1}\big)\operatorname{vec}\big(\mathcal{Y}_t-\mathcal{B}\bar{\times}_1\mathbf{x}_t\big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\sum_{t=1}^{T}\operatorname{vec}\big(\tilde{\mathcal{E}}_t\big)'\big(\Sigma_3^{-1}\otimes\Sigma_2^{-1}\otimes\Sigma_1^{-1}\big)\operatorname{vec}\big(\tilde{\mathcal{E}}_t\big)\Big).
\end{aligned}
$$

Thus, defining with $\mathbf{y}_t=\operatorname{vec}\big(\mathcal{Y}_t\big)$ and $\mathbf{\Sigma}^{-1}=\Sigma_3^{-1}\otimes\Sigma_2^{-1}\otimes\Sigma_1^{-1}$, we obtain

$$
\begin{aligned}
&L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3)\propto \\
&\propto \exp\Big(-\frac{1}{2}\sum_{t=1}^{T}\operatorname{vec}\big(\tilde{\mathcal{E}}_t\big)'\big(\Sigma_3^{-1}\otimes\Sigma_2^{-1}\otimes\Sigma_1^{-1}\big)\operatorname{vec}\big(\tilde{\mathcal{E}}_t\big)\Big) \\
&\propto \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T}\Big(\operatorname{vec}\big(\mathcal{Y}_t\big)-\operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)-\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big) \\
&\quad\cdot\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\Big)'\mathbf{\Sigma}^{-1}\Big(\operatorname{vec}\big(\mathcal{Y}_t\big)-\operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)-\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\Big)\Bigg) \\
&= \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T}\mathbf{y}_t'\mathbf{\Sigma}^{-1}\mathbf{y}_t-2\mathbf{y}_t'\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big) \\
&\quad+\operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big) \\
&\quad-2\mathbf{y}_t'\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle \\
&\quad+2\operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)'\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle \\
&\quad+\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)'\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T}-2\big(\mathbf{y}_t'-\operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle \\
&\quad+\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)'\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\Bigg).
\end{aligned}
\tag{S22}
$$

Now, we focus on a specific $j=1,2,3,4$ and derive proportionality results that will be necessary to obtain the posterior full conditional distributions of the PARAFAC marginals of the tensor $\mathcal{B}$. Consider the case $j=1$. By exploiting eq. (S19) we get

$$
L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3)\propto
$$

$$\propto \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T} -2\big(\mathbf{y}_t' - \mathrm{vec}\,(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t)'\big)\boldsymbol{\Sigma}^{-1}\,\mathrm{vec}\,(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)})\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)}$$

$$+\,\mathrm{vec}\,(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)})'\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\boldsymbol{\Sigma}^{-1}\,\mathrm{vec}\,(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)})\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T} -2\big(\mathbf{y}_t' - \mathrm{vec}\,(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t)'\big)\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)\boldsymbol{\beta}_1^{(r)}$$

$$+\,\big(\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)\boldsymbol{\beta}_1^{(r)}\big)'\boldsymbol{\Sigma}^{-1}\big(\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)\boldsymbol{\beta}_1^{(r)}\big)\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\beta}_1^{(r)\prime}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)'\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)\boldsymbol{\beta}_1^{(r)}$$

$$-\,2\big(\mathbf{y}_t' - \mathrm{vec}\,(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t)'\big)\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)\boldsymbol{\beta}_1^{(r)}\Bigg)$$

$$= \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_1^{(r)\prime}\mathbf{S}_1^L\boldsymbol{\beta}_1^{(r)} - 2\mathbf{m}_1^L\boldsymbol{\beta}_1^{(r)}\Big), \tag{S23}$$

with

$$\mathbf{S}_1^L = \sum_{t=1}^{T}\big(\boldsymbol{\beta}_3^{(r)\prime}\otimes\boldsymbol{\beta}_2^{(r)\prime}\otimes\mathbf{I}_{I_1}\big)\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2$$

$$\mathbf{m}_1^L = \sum_{t=1}^{T}\big(\mathbf{y}_t' - \mathrm{vec}\,(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t)'\big)\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle.$$

Consider the case $j=2$. From eq. (S20) we get

$$L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3)\propto$$

$$\propto \exp\Big(-\frac{1}{2}\sum_{t=1}^{T} -2\big(\mathbf{y}_t' - \mathrm{vec}\,(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t)'\big)\boldsymbol{\Sigma}^{-1}\,\mathrm{vec}\,(\boldsymbol{\beta}_1^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)})\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)}$$

$$+\,\mathrm{vec}\,(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)})'\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\boldsymbol{\Sigma}^{-1}\,\mathrm{vec}\,(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)})\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T} -2\big(\mathbf{y}_t' - \mathrm{vec}\,(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t)'\big)\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\circ\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_2^{(r)}$$

$$+\,\big(\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_2^{(r)}\big)'\boldsymbol{\Sigma}^{-1}\big(\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_2^{(r)}\big)\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\beta}_2^{(r)\prime}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_2^{(r)}$$

$$
\begin{aligned}
&\quad - 2\big(\mathbf{y}_t' - \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_2^{(r)}\Bigg) \\
&= \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_2^{(r)\prime}\mathbf{S}_2^L\boldsymbol{\beta}_2^{(r)} - 2\mathbf{m}_2^L\boldsymbol{\beta}_2^{(r)}\Big),
\end{aligned} \tag{S24}
$$

with

$$
\begin{aligned}
\mathbf{S}_2^L &= \sum_{t=1}^T \big(\boldsymbol{\beta}_3^{(r)\prime}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)\prime}\big)\mathbf{\Sigma}^{-1}\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2 \\
\mathbf{m}_2^L &= \sum_{t=1}^T \big(\mathbf{y}_t' - \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\big(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle.
\end{aligned}
$$

Consider the case $j = 3$, by exploiting eq. (S21) we get

$$
\begin{aligned}
&L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3) \propto \\
&\propto \exp\Bigg(-\frac{1}{2}\sum_{t=1}^T -2\big(\mathbf{y}_t' - \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} \\
&\quad + \big(\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big)'\mathbf{\Sigma}^{-1}\big(\operatorname{vec}\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big)\Bigg) \\
&= \exp\Bigg(-\frac{1}{2}\sum_{t=1}^T -2\big(\mathbf{y}_t' - \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_3^{(r)} \\
&\quad + \big(\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_3^{(r)}\big)'\mathbf{\Sigma}^{-1}\big(\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_3^{(r)}\big)\Bigg) \\
&= \exp\Bigg(-\frac{1}{2}\sum_{t=1}^T \boldsymbol{\beta}_3^{(r)\prime}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\mathbf{\Sigma}^{-1}\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_3^{(r)} \\
&\quad - 2\big(\mathbf{y}_t' - \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\boldsymbol{\beta}_3^{(r)}\Bigg) \\
&= \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_3^{(r)\prime}\mathbf{S}_3^L\boldsymbol{\beta}_3^{(r)} - 2\mathbf{m}_3^L\boldsymbol{\beta}_3^{(r)}\Big),
\end{aligned} \tag{S25}
$$

with

$$
\begin{aligned}
\mathbf{S}_3^L &= \sum_{t=1}^T \big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)\prime}\otimes\boldsymbol{\beta}_1^{(r)\prime}\big)\mathbf{\Sigma}^{-1}\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2 \\
\mathbf{m}_3^L &= \sum_{t=1}^T \big(\mathbf{y}_t' - \operatorname{vec}\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\big(\mathbf{I}_{I_3}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\boldsymbol{\beta}_1^{(r)}\big)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle.
\end{aligned}
$$

Finally, in the case $j = 4$. From eq. (S22) we get

$$
L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3) \propto
$$

$$
\begin{aligned}
&\propto \exp\Bigg(-\frac{1}{2}\sum_{t=1}^{T} -2\big(\mathbf{y}_t' - \mathrm{vec}\,\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\,\mathrm{vec}\,\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} \\
&\quad + \boldsymbol{\beta}_4^{(r)\prime}\mathbf{x}_t\,\mathrm{vec}\,\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)'\mathbf{\Sigma}^{-1}\,\mathrm{vec}\,\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)}\Bigg) \\
&= \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_4^{(r)\prime}\mathbf{S}_4^L\boldsymbol{\beta}_4^{(r)} - 2\mathbf{m}_4^L\boldsymbol{\beta}_4^{(r)}\Big),
\end{aligned} \tag{S26}
$$

with

$$
\begin{aligned}
\mathbf{S}_4^L &= \sum_{t=1}^{T}\mathbf{x}_t\,\mathrm{vec}\,\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)'\mathbf{\Sigma}^{-1}\,\mathrm{vec}\,\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\mathbf{x}_t' \\
\mathbf{m}_4^L &= \sum_{t=1}^{T}\big(\mathbf{y}_t' - \mathrm{vec}\,\big(\mathcal{B}_{-r}\bar{\times}_1\mathbf{x}_t\big)'\big)\mathbf{\Sigma}^{-1}\,\mathrm{vec}\,\big(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\big)\mathbf{x}_t'.
\end{aligned}
$$

It is now possible to derive the full conditional distributions for the PARAFAC marginals $\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)}, \boldsymbol{\beta}_4^{(r)}$, as shown in the following.

S.6.5.1 Full conditional distribution of $\beta_1^{(r)}$

The posterior full conditional distribution of $\boldsymbol{\beta}_1^{(r)}$ is obtained by combining the prior distribution in eq. (17) and the likelihood in eq. (S23) as follows

$$
\begin{aligned}
&p(\boldsymbol{\beta}_1^{(r)}|\boldsymbol{\beta}_{-1}^{(r)},\mathcal{B}_{-r},W_{1,r},\phi_r,\tau,\Sigma_1,\Sigma_2,\Sigma_3,\mathbf{Y}) \propto L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3)\pi(\boldsymbol{\beta}_1^{(r)}|W_{1,r},\phi_r,\tau) \\
&\propto \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_1^{(r)\prime}\mathbf{S}_1^L\boldsymbol{\beta}_1^{(r)} - 2\mathbf{m}_1^L\boldsymbol{\beta}_1^{(r)}\Big)\cdot\exp\Big(-\frac{1}{2}\boldsymbol{\beta}_1^{(r)\prime}(W_{1,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_1^{(r)}\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_1^{(r)\prime}\mathbf{S}_1^L\boldsymbol{\beta}_1^{(r)} - 2\mathbf{m}_1^L\boldsymbol{\beta}_1^{(r)} + \boldsymbol{\beta}_1^{(r)\prime}(W_{1,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_1^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_1^{(r)\prime}\big(\mathbf{S}_1^L + (W_{1,r}\phi_r\tau)^{-1}\big)\boldsymbol{\beta}_1^{(r)} - 2\mathbf{m}_1^L\boldsymbol{\beta}_1^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_1^{(r)\prime}\bar{\Sigma}_{\boldsymbol{\beta}_1^r}^{-1}\boldsymbol{\beta}_1^{(r)} - 2\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_1^r}\boldsymbol{\beta}_1^{(r)}\big)\Big),
\end{aligned}
$$

where

$$
\bar{\Sigma}_{\boldsymbol{\beta}_1^r} = \big((W_{1,r}\phi_r\tau)^{-1} + \mathbf{S}_1^L\big)^{-1}, \qquad \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_1^r} = \bar{\Sigma}_{\boldsymbol{\beta}_1^r}(\mathbf{m}_1^L)'.
$$

Thus the posterior full conditional distribution of $\boldsymbol{\beta}_1^{(r)}$ is given by

$$
p(\boldsymbol{\beta}_1^{(r)}|\boldsymbol{\beta}_{-1}^{(r)},\mathcal{B}_{-r},W_{1,r},\phi_r,\tau,\Sigma_1,\Sigma_2,\Sigma_3,\mathbf{Y}) \sim \mathcal{N}_{I_1}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_1^r},\bar{\Sigma}_{\boldsymbol{\beta}_1^r}).
$$

182 ### S.6.5.2 Full conditional distribution of $\beta_2^{(r)}$

The posterior full conditional distribution of $\boldsymbol{\beta}_2^{(r)}$ is obtained by combining the prior distribution in eq. (17) and the likelihood in eq. (S24) as follows

$$
\begin{aligned}
&p(\boldsymbol{\beta}_2^{(r)}|\boldsymbol{\beta}_{-2}^{(r)}, \mathcal{B}_{-r}, W_{2,r}, \phi_r, \tau, \Sigma_1, \Sigma_2, \Sigma_3, \mathbf{Y}) \propto L(\mathbf{Y}|\mathcal{B}, \Sigma_1, \Sigma_2, \Sigma_3)\pi(\boldsymbol{\beta}_2^{(r)}|W_{2,r}, \phi_r, \tau) \\
&\propto \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_2^{(r)\prime}\mathbf{S}_2^L\boldsymbol{\beta}_2^{(r)} - 2\mathbf{m}_2^L\boldsymbol{\beta}_2^{(r)}\Big) \cdot \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_2^{(r)\prime}(W_{2,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_2^{(r)}\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_2^{(r)\prime}\mathbf{S}_2^L\boldsymbol{\beta}_2^{(r)} - 2\mathbf{m}_2^L\boldsymbol{\beta}_2^{(r)} + \boldsymbol{\beta}_2^{(r)\prime}(W_{2,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_2^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_2^{(r)\prime}\big(\mathbf{S}_2^L + (W_{2,r}\phi_r\tau)^{-1}\big)\boldsymbol{\beta}_2^{(r)} - 2\mathbf{m}_2^L\boldsymbol{\beta}_2^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_2^{(r)\prime}\bar{\Sigma}_{\boldsymbol{\beta}_2^r}^{-1}\boldsymbol{\beta}_2^{(r)} - 2\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_2^r}\boldsymbol{\beta}_2^{(r)}\big)\Big),
\end{aligned}
$$

where

$$
\bar{\Sigma}_{\boldsymbol{\beta}_2^r} = \big((W_{2,r}\phi_r\tau)^{-1} + \mathbf{S}_2^L\big)^{-1}, \qquad \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_2^r} = \bar{\Sigma}_{\boldsymbol{\beta}_2^r}(\mathbf{m}_2^L)'.
$$

Thus the posterior full conditional distribution of $\boldsymbol{\beta}_2^{(r)}$ is given by

$$
p(\boldsymbol{\beta}_2^{(r)}|\boldsymbol{\beta}_{-2}^{(r)}, \mathcal{B}_{-r}, W_{2,r}, \phi_r, \tau, \Sigma_1, \Sigma_2, \Sigma_3, \mathbf{Y}) \sim \mathcal{N}_{I_2}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_2^r}, \bar{\Sigma}_{\boldsymbol{\beta}_2^r}).
$$

183 ### S.6.5.3 Full conditional distribution of $\beta_3^{(r)}$

The posterior full conditional distribution of $\boldsymbol{\beta}_3^{(r)}$ is obtained by combining the prior distribution in eq. (17) and the likelihood in eq. (S25) as follows

$$
\begin{aligned}
&p(\boldsymbol{\beta}_3^{(r)}|\boldsymbol{\beta}_{-3}^{(r)}, \mathcal{B}_{-r}, W_{3,r}, \phi_r, \tau, \Sigma_1, \Sigma_2, \Sigma_3, \mathbf{Y}) \propto L(\mathbf{Y}|\mathcal{B}, \Sigma_1, \Sigma_2, \Sigma_3)\pi(\boldsymbol{\beta}_3^{(r)}|W_{3,r}, \phi_r, \tau) \\
&\propto \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_3^{(r)\prime}\mathbf{S}_3^L\boldsymbol{\beta}_3^{(r)} - 2\mathbf{m}_3^L\boldsymbol{\beta}_3^{(r)}\Big) \cdot \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_3^{(r)\prime}(W_{3,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_3^{(r)}\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_3^{(r)\prime}\mathbf{S}_3^L\boldsymbol{\beta}_3^{(r)} - 2\mathbf{m}_3^L\boldsymbol{\beta}_3^{(r)} + \boldsymbol{\beta}_3^{(r)\prime}(W_{3,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_3^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_3^{(r)\prime}\big(\mathbf{S}_3^L + (W_{3,r}\phi_r\tau)^{-1}\big)\boldsymbol{\beta}_3^{(r)} - 2\mathbf{m}_3^L\boldsymbol{\beta}_3^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_3^{(r)\prime}\bar{\Sigma}_{\boldsymbol{\beta}_3^r}^{-1}\boldsymbol{\beta}_3^{(r)} - 2\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_3^r}\boldsymbol{\beta}_3^{(r)}\big)\Big),
\end{aligned}
$$

where

$$
\bar{\Sigma}_{\boldsymbol{\beta}_3^r} = \big((W_{3,r}\phi_r\tau)^{-1} + \mathbf{S}_3^L\big)^{-1}, \qquad \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_3^r} = \bar{\Sigma}_{\boldsymbol{\beta}_3^r}(\mathbf{m}_3^L)'.
$$

Thus the posterior full conditional distribution of $\boldsymbol{\beta}_3^{(r)}$ is given by

$$
p(\boldsymbol{\beta}_3^{(r)}|\boldsymbol{\beta}_{-3}^{(r)}, \mathcal{B}_{-r}, W_{3,r}, \phi_r, \tau, \Sigma_1, \Sigma_2, \Sigma_3, \mathbf{Y}) \sim \mathcal{N}_{I_3}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_3^r}, \bar{\Sigma}_{\boldsymbol{\beta}_3^r}).
$$

184 S.6.5.4 Full conditional distribution of $\beta_4^{(r)}$

The posterior full conditional distribution of $\boldsymbol{\beta}_4^{(r)}$ is obtained by combining the prior distribution in eq. (17) and the likelihood in eq. (S26) as follows

$$
\begin{aligned}
&p(\boldsymbol{\beta}_4^{(r)}|\boldsymbol{\beta}_{-4}^{(r)}, \mathcal{B}_{-r}, W_{4,r}, \phi_r, \tau, \Sigma_1, \Sigma_2, \Sigma_3, \mathbf{Y}) \propto L(\mathbf{Y}|\mathcal{B}, \Sigma_1, \Sigma_2, \Sigma_3)\pi(\boldsymbol{\beta}_4^{(r)}|W_{4,r}, \phi_r, \tau) \\
&\propto \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_4^{(r)\prime}\mathbf{S}_4^L\boldsymbol{\beta}_4^{(r)} - 2\mathbf{m}_4^L\boldsymbol{\beta}_4^{(r)}\Big) \cdot \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_4^{(r)\prime}(W_{4,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_4^{(r)}\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_4^{(r)\prime}\mathbf{S}_4^L\boldsymbol{\beta}_4^{(r)} - 2\mathbf{m}_4^L\boldsymbol{\beta}_4^{(r)} + \boldsymbol{\beta}_4^{(r)\prime}(W_{4,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_4^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_4^{(r)\prime}\big(\mathbf{S}_4^L + (W_{4,r}\phi_r\tau)^{-1}\big)\boldsymbol{\beta}_4^{(r)} - 2\mathbf{m}_4^L\boldsymbol{\beta}_4^{(r)}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\boldsymbol{\beta}_4^{(r)\prime}\bar{\Sigma}_{\boldsymbol{\beta}_4^r}^{-1}\boldsymbol{\beta}_4^{(r)} - 2\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_4^r}\boldsymbol{\beta}_4^{(r)}\big)\Big),
\end{aligned}
$$

where

$$
\bar{\Sigma}_{\boldsymbol{\beta}_4^r} = \big((W_{4,r}\phi_r\tau)^{-1} + \mathbf{S}_4^L\big)^{-1}, \qquad \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_4^r} = \bar{\Sigma}_{\boldsymbol{\beta}_4^r}(\mathbf{m}_4^L)'.
$$

Thus the posterior full conditional distribution of $\boldsymbol{\beta}_4^{(r)}$ is given by

$$
p(\boldsymbol{\beta}_4^{(r)}|\boldsymbol{\beta}_{-4}^{(r)}, \mathcal{B}_{-r}, W_{4,r}, \phi_r, \tau, \Sigma_1, \Sigma_2, \Sigma_3, \mathbf{Y}) \sim \mathcal{N}_{I_1 I_2 I_3}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_4^r}, \bar{\Sigma}_{\boldsymbol{\beta}_4^r}).
$$

185 S.6.6 Full conditional distribution of Σ_1

Given a inverse Wishart prior, the posterior full conditional distribution for Σ_1 is conjugate. For ease of notation, define $\tilde{\mathcal{E}}_t = \mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t$, $\tilde{\mathbf{E}}_{(1),t}$ the mode-1 matricization of $\tilde{\mathcal{E}}_t$ and $\mathbf{Z}_1 = \Sigma_3^{-1}\otimes\Sigma_2^{-1}$. By exploiting the relation between the tensor normal distribution and the multivariate normal distribution and the properties of the vectorization and trace operators, we obtain

$$
\begin{aligned}
&p(\Sigma_1|\mathcal{B}, \mathbf{Y}, \Sigma_2, \Sigma_3, \gamma) \propto L(\mathbf{Y}|\mathcal{B}, \Sigma_1, \Sigma_2, \Sigma_3)\pi(\Sigma_1|\gamma) \\
&\propto |\Sigma_1|^{-\frac{TI_2I_3}{2}} \exp\Big(-\frac{1}{2}\sum_{t=1}^{T} \mathrm{vec}\left(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t\right)'(\Sigma_3^{-1}\otimes\Sigma_2^{-1}\otimes\Sigma_1^{-1}) \\
&\quad \cdot \mathrm{vec}\left(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t\right)\Big) \cdot |\Sigma_1|^{-\frac{\nu_1+I_1+1}{2}} \exp\Big(-\frac{1}{2}\mathrm{tr}\left(\gamma\Psi_1\Sigma_1^{-1}\right)\Big) \\
&\propto |\Sigma_1|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\Big(-\frac{1}{2}\Big(\mathrm{tr}\left(\gamma\Psi_1\Sigma_1^{-1}\right) + \sum_{t=1}^{T}\mathrm{vec}\left(\tilde{\mathcal{E}}_t\right)'(\mathbf{Z}_1\otimes\Sigma_1^{-1})\,\mathrm{vec}\left(\tilde{\mathcal{E}}_t\right)\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&\propto \left|\Sigma_1\right|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\left(\gamma\Psi_1\Sigma_1^{-1}\right) \\
&\quad + \sum_{t=1}^{T} \operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)'(\mathbf{Z}_1 \otimes \Sigma_1^{-1}) \operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)\big)\Big) \\
&\propto \left|\Sigma_1\right|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\left(\gamma\Psi_1\Sigma_1^{-1}\right) \\
&\quad + \sum_{t=1}^{T} \operatorname{tr}\big(\operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)' \operatorname{vec}\left(\Sigma_1^{-1}\tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\right)\big)\big)\Big) \\
&\propto \left|\Sigma_1\right|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\left(\gamma\Psi_1\Sigma_1^{-1}\right) + \sum_{t=1}^{T} \operatorname{tr}\left(\tilde{\mathbf{E}}'_{(1),t}\Sigma_1^{-1}\tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\right)\big)\Big) \\
&\propto \left|\Sigma_1\right|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\left(\gamma\Psi_1\Sigma_1^{-1}\right) + \sum_{t=1}^{T} \operatorname{tr}\left(\tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\tilde{\mathbf{E}}'_{(1),t}\Sigma_1^{-1}\right)\big)\Big).
\end{aligned}
$$

For ease of notation, define $S_1 = \sum_{t=1}^{T} \tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\tilde{\mathbf{E}}'_{(1),t}$. Then

$$
\begin{aligned}
p(\Sigma_1|\mathcal{B},\mathbf{Y},\Sigma_2,\Sigma_3) &\propto \left|\Sigma_1\right|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\left(\gamma\Psi_1\Sigma_1^{-1}\right) + \operatorname{tr}\left(S_1\Sigma_1^{-1}\right)\big)\Big) \\
&\propto \left|\Sigma_1\right|^{-\frac{(\nu_1+TI_2I_3)+I_1+1}{2}} \exp\Big(-\frac{1}{2} \operatorname{tr}\left((\gamma\Psi_1 + S_1)\Sigma_1^{-1}\right)\Big),
\end{aligned}
$$

Therefore, the posterior full conditional distribution of Σ_1 is given by

$$
p(\Sigma_1|\mathcal{B},\mathbf{Y},\Sigma_2,\Sigma_3,\gamma) \sim \mathcal{IW}_{I_1}\big(\nu_1 + TI_2I_3, \gamma\Psi_1 + S_1\big).
$$

186 S.6.7 Full conditional distribution of Σ_2

Given a inverse Wishart prior, the posterior full conditional distribution for Σ_2 is conjugate. For ease of notation, define $\tilde{\mathcal{E}}_t = \mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t$ and $\tilde{\mathbf{E}}_{(2),t}$ the mode-2 matricization of $\tilde{\mathcal{E}}_t$. By exploiting the relation between the tensor normal distribution and the matrix normal distribution and the properties of the Kronecker product and of the vectorization and trace operators we obtain

$$
\begin{aligned}
&p(\Sigma_2|\mathcal{B},\mathbf{Y},\Sigma_1,\Sigma_3,\gamma) \propto L(\mathbf{Y}|\mathcal{B},\Sigma_1,\Sigma_2,\Sigma_3)\pi(\Sigma_2|\gamma) \\
&\propto \left|\Sigma_2\right|^{-\frac{TI_1I_3}{2}} \exp\Big(-\frac{1}{2}\sum_{t=1}^{T}(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t)\bar{\times}_3(\Sigma_1^{-1}\circ\Sigma_2^{-1}\circ\Sigma_3^{-1}) \\
&\quad \bar{\times}_3(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t)\Big) \cdot \left|\Sigma_2\right|^{-\frac{\nu_2+I_2+1}{2}} \exp\Big(-\frac{1}{2} \operatorname{tr}\left(\Psi_2\Sigma_2^{-1}\right)\Big) \\
&\propto \left|\Sigma_2\right|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\left(\gamma\Psi_2\Sigma_2^{-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{t=1}^{T} \tilde{\mathcal{E}}_t \bar{\times}_3 (\Sigma_1^{-1} \circ \Sigma_2^{-1} \circ \Sigma_3^{-1}) \bar{\times}_3 \tilde{\mathcal{E}}_t)\Big) \\
\propto & \big|\Sigma_2\big|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_2\Sigma_2^{-1}\big) \\
&+\sum_{t=1}^{T} \operatorname{tr}\big(\tilde{\mathbf{E}}'_{(2),t}(\Sigma_3^{-1} \otimes \Sigma_1^{-1} \otimes \Sigma_2^{-1})\tilde{\mathbf{E}}_{(2),t})\big)\Big) \\
\propto & \big|\Sigma_2\big|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_2\Sigma_2^{-1}\big) \\
&+\sum_{t=1}^{T} \operatorname{tr}\big((\Sigma_3^{-1} \otimes \Sigma_1^{-1})\tilde{\mathbf{E}}'_{(2),t}\Sigma_2^{-1}\tilde{\mathbf{E}}_{(2),t})\big)\Big) \\
\propto & \big|\Sigma_2\big|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_2\Sigma_2^{-1}\big) + \operatorname{tr}\big(\sum_{t=1}^{T}\tilde{\mathbf{E}}_{(2),t}(\Sigma_3^{-1} \otimes \Sigma_1^{-1})\tilde{\mathbf{E}}'_{(2),t}\Sigma_2^{-1})\big)\Big) \\
\propto & \big|\Sigma_2\big|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\Big(-\frac{1}{2}\operatorname{tr}\big(\gamma\Psi_2\Sigma_2^{-1} + S_2\Sigma_2^{-1}\big)\Big),
\end{aligned}
$$

where for ease of notation we defined $S_2 = \sum_{t=1}^{T} \tilde{\mathbf{E}}_{(2),t}(\Sigma_3^{-1} \otimes \Sigma_1^{-1})\tilde{\mathbf{E}}'_{(2),t}$. Therefore, the posterior full conditional distribution of Σ_2 is given by

$$
p(\Sigma_2|\mathcal{B}, \mathbf{Y}, \Sigma_1, \Sigma_3) \sim \mathcal{IW}_{I_2}(\nu_2 + TI_1I_3, \gamma\Psi_2 + S_2).
$$

187 ### S.6.8 Full conditional distribution of Σ_3

Given a inverse Wishart prior, the posterior full conditional distribution for Σ_3 is conjugate. For ease of notation, define $\tilde{\mathcal{E}}_t = \mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t$, $\tilde{\mathbf{E}}_{(3),t}$ the mode-3 matricization of $\tilde{\mathcal{E}}_t$ and $\mathbf{Z}_3 = \Sigma_2^{-1} \otimes \Sigma_1^{-1}$. By exploiting the relation between the tensor normal distribution and the multivariate normal distribution and the properties of the vectorization and trace operators, we obtain

$$
\begin{aligned}
&p(\Sigma_3|\mathcal{B}, \mathbf{Y}, \Sigma_1, \Sigma_2, \gamma) \propto L(\mathbf{Y}|\mathcal{B}, \Sigma_1, \Sigma_2, \Sigma_3)\pi(\Sigma_3|\gamma) \\
&\propto \big|\Sigma_3\big|^{-\frac{TI_1I_2}{2}} \exp\Big(-\frac{1}{2}\sum_{t=1}^{T} \operatorname{vec}\big(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t\big)'(\Sigma_3^{-1} \otimes \Sigma_2^{-1} \otimes \Sigma_1^{-1}) \\
&\quad \cdot \operatorname{vec}\big(\mathcal{Y}_t - \mathcal{B}\bar{\times}_1\mathbf{x}_t\big)\Big) \cdot \big|\Sigma_3\big|^{-\frac{\nu_3+I_3+1}{2}} \exp\Big(-\frac{1}{2}\operatorname{tr}\big(\gamma\Psi_3\Sigma_3^{-1}\big)\Big) \\
&\propto \big|\Sigma_3\big|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_3\Sigma_3^{-1}\big) + \sum_{t=1}^{T} \operatorname{vec}\big(\tilde{\mathcal{E}}_t\big)'(\Sigma_3^{-1} \otimes \mathbf{Z}_3)\operatorname{vec}\big(\tilde{\mathcal{E}}_t\big)\big)\Big) \\
&\propto \big|\Sigma_3\big|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_3\Sigma_3^{-1}\big)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{t=1}^{T} \operatorname{vec}\big(\tilde{\mathbf{E}}_{(3),t}\big)'(\Sigma_3^{-1}\otimes \mathbf{Z}_3)\operatorname{vec}\big(\tilde{\mathbf{E}}_{(3),t}\big)\big)\Big)\\
\propto&\ \big|\Sigma_3\big|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_3\Sigma_3^{-1}\big)\\
&+\sum_{t=1}^{T} \operatorname{tr}\big(\operatorname{vec}\big(\tilde{\mathbf{E}}_{(3),t}\big)'\operatorname{vec}\big(\mathbf{Z}_3\tilde{\mathbf{E}}_{(3),t}\Sigma_3^{-1}\big)\big)\big)\Big)\\
\propto&\ \big|\Sigma_3\big|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_3\Sigma_3^{-1}\big)+\sum_{t=1}^{T}\operatorname{tr}\big(\tilde{\mathbf{E}}'_{(3),t}\mathbf{Z}_3\tilde{\mathbf{E}}_{(3),t}\Sigma_3^{-1}\big)\big)\Big).
\end{aligned}
$$

For ease of notation, define $S_3=\sum_{t=1}^{T}\tilde{\mathbf{E}}_{(3),t}\mathbf{Z}_3\tilde{\mathbf{E}}'_{(3),t}$. Then

$$
\begin{aligned}
p(\Sigma_3|\mathcal{B},\mathbf{Y},\Sigma_1,\Sigma_2) &\propto \big|\Sigma_3\big|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\Big(-\frac{1}{2}\big(\operatorname{tr}\big(\gamma\Psi_3\Sigma_3^{-1}\big)+\operatorname{tr}\big(S_3\Sigma_3^{-1}\big)\big)\Big)\\
&\propto \big|\Sigma_3\big|^{-\frac{(\nu_3+TI_1I_2)+I_3+1}{2}} \exp\Big(-\frac{1}{2}\operatorname{tr}\big((\gamma\Psi_3+S_3)\Sigma_3^{-1}\big)\Big),
\end{aligned}
$$

Therefore, the posterior full conditional distribution of Σ_3 is given by

$$
p(\Sigma_3|\mathcal{B},\mathbf{Y},\Sigma_1,\Sigma_2)\sim \mathcal{IW}_{I_3}(\nu_3+TI_1I_2,\gamma\Psi_3+S_3).
$$

S.6.9 Full conditional distribution of γ

Using a gamma prior distribution we have

$$
\begin{aligned}
p(\gamma|\Sigma_1,\Sigma_2,\Sigma_3) &\propto p(\Sigma_1,\Sigma_2,\Sigma_3|\gamma)\pi(\gamma)\\
&\propto \prod_{i=1}^{3}\big|\gamma\Psi_i\big|^{-\frac{\nu_i}{2}}\exp\Big(-\frac{1}{2}\operatorname{tr}\big(\gamma\Psi_i\Sigma_i^{-1}\big)\Big)\gamma^{a_\gamma-1}e^{-b_\gamma\gamma}\\
&\propto \gamma^{a_\gamma-\frac{\sum_{i=1}^{3}\nu_iI_i}{2}-1}\exp\Big(-\frac{1}{2}\operatorname{tr}\big(\sum_{i=1}^{3}\Psi_i\Sigma_i^{-1}\big)-b_\gamma\gamma\Big),
\end{aligned}
$$

thus

$$
p(\gamma|\Sigma_1,\Sigma_2,\Sigma_3)\sim \mathcal{G}a\Big(a_\gamma+\frac{1}{2}\sum_{i=1}^{3}\nu_iI_i,b_\gamma+\frac{1}{2}\operatorname{tr}\big(\sum_{i=1}^{3}\Psi_i\Sigma_i^{-1}\big)\Big).
$$

S.7 Initialisation details

It is well known that the Gibbs sampler algorithm is highly sensitive to the choice of the initial value. From this point of view, the most difficult parameters initialise in the proposed model are the margins of the tensor of coefficients, that is the set of vectors: $(\boldsymbol{\beta}_1^{(r)}, \dots, \boldsymbol{\beta}_J^{(r)})_{r=1}^R$. Due to the high complexity of the parameter space, we have chosen to perform an initialisation scheme which is based on the Simulated Annealing (SA) algorithm (see Press et al., 2007; Robert and Casella, 2004). This algorithm is similar to the Metropolis-Hastings one, and the idea behind it is to perform a stochastic optimisation by proposing random moves from the current state which are always accepted when improving the optimum and have positive probability of acceptance even when they are not improving. This is used in order to allow the algorithm to escape from local optima. Denoting the objective function to be minimised by $f(\boldsymbol{\theta})$, the Simulated Annealing method accepts a move from the current state $\boldsymbol{\theta}^{(i)}$ to the proposed one $\boldsymbol{\theta}^*$ with probability given by the Bolzmann-like distribution

$$p(\Delta f, T) = \exp\Big(-\frac{\Delta f}{T} \Big).$$

Here $\Delta f = f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^{(i)})$ and T is a parameter called temperature. The key of the SA method is in the cooling scheme, which describes the deterministic, decreasing evolution of the temperature over the iterations of the algorithm: it has been proved that under sufficiently slow decreasing schemes, the SA yields a global optimum.

We use the SA algorithm for minimising the objective function

$$f((\boldsymbol{\beta}_j^{(r)})_{j,r}) = \kappa_N \psi_N + \kappa_J \psi_J,$$

where κ_N is an overall penalty given by the Frobenius norm of the tensor constructed from simulated margins, while κ_J is the penalty of the sum (over r) of the norms of the marginals $\boldsymbol{\beta}_J^{(r)}$. In formulas:

$$\psi_N = \left\| \boldsymbol{\mathcal{B}}^{SA} \right\|_2 \qquad \psi_J = \sum_{r=1}^{R} \left\| \boldsymbol{\beta}_J^{(r)} \right\|_2.$$

The proposal distribution for each margin is a normal $\mathcal{N}_{I_j}(\mathbf{0}, \sigma \mathbf{I}_{I_j})$, independent from the current state of the algorithm. Finally, we have chosen a logarithmic cooling scheme which updates the temperature at each iteration of the SA

$$T_i = \frac{k}{1 + \log(i)} \qquad i = 1, \dots, I^{SA},$$

where $k > 0$ is a tuning parameter, which can be interpreted as the initial value of the temperature. In order to perform the initialisation of the margins, we run the SA algorithm for $I^{SA} = 1,200$ iterations, then we took the vectors which gave the best fit in terms of minimum value of the objective function.

S.8 Simulation Results

We report the results of a simulation study where we have tested the performance of the proposed sampler on synthetic datasets of matrix-valued sequences $(Y_t, X_t)_{t=1}^T$, where Y_t, X_t have different size across simulations. The methods described in this paper can be rather computationally intensive, nevertheless thanks to the tensor decomposition we used allows the estimation to be carried out on a laptop. All the simulations were run on an Apple MacBookPro with a 3.1GHz Intel Core i7 processor, RAM 16GB, using MATLAB r2017b with the aid of the Tensor Toolbox v.2.6[2].

We have fixed $I_1 = I_2 = I$ and performed experiments for different sizes I of the response and covariate matrices. We have generated a matrix-valued time series $(Y_t, X_t)_{t=1}^T$ by simulating each entry of X_t from

$$x_{ij,t} - \mu = \alpha_{ij}(x_{ij,t-1} - \mu) + \eta_{ij,t}, \qquad \eta_{ij,t} \sim \mathcal{N}(0,1),$$

for $i = 1, \dots, I_1$, $j = 1, \dots, I_2$ and $t = 1, \dots, T$. Then, we have generated the matrix-valued response Y_t according to

$$Y_t = \mathcal{B} \bar{\times}_1 \operatorname{vec}\big(X_t\big) + E_t, \qquad E_t \sim \mathcal{N}_{I_1,I_2}(\mathbf{0}, \Sigma_1, \mathbf{I}_{I_2}).$$

where $\mathbb{E}(\eta_{ij,t}\eta_{kl,v}) = 0$, $\mathbb{E}(\eta_{ij,t}E_v) = 0$, $\forall\,(i,j) \neq (k,l)$, $\forall\, t \neq v$, and $\alpha_{ij} \sim \mathcal{U}(-1,1)$. We randomly draw $\mathcal{B}$ using the PARAFAC decomposition in eq. (1), with rank $R = 5$ and marginals sampled from the prior distribution in eq. (17). The matrices X_t, Y_t in each simulated dataset have size $I \in \{10, 20, 30, 40\}$, and $T = 60$ in each simulation. We initialized the Gibbs sampler by setting the PARAFAC marginals $\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)}$, $r = 1, \dots, R$ (with $R = 5$), via simulated annealing (see section S.7). We chose a burn-in period of $10,000$ iterations and, due to autocorrelation in the sample, we applied thinning and selected every 2nd iteration, thus obtaining $5,000$ draws from the posterior distribution after convergence.

[2] http://www.sandia.gov/~tgkolda/TensorToolbox/index-2.6.html

Fig. 5 shows the accuracy of the sampler in estimating the coefficient tensor, in the four experiments corresponding to $I \in \{10, 20, 30, 40\}$. The efficiency decreases with I (recall that the number of cells of the coefficient tensor is I^4). The estimation error is mainly due to the over-shrinking to zero, which is a known drawback of global-local hierarchical prior distributions (e.g, see Carvalho et al., 2010). Note that we expected a decrease of efficiency with I, since the sample size was held fixed ($T = 60$) across all simulation experiments, while increasing the size of the parameter space. In Figg. 6, 8, 10, 12 we report the estimation results for some randomly chosen cells of the coefficient tensor. We find that, after removing burn-in iterations and performing thinnig, the autocorrelation wipes out.

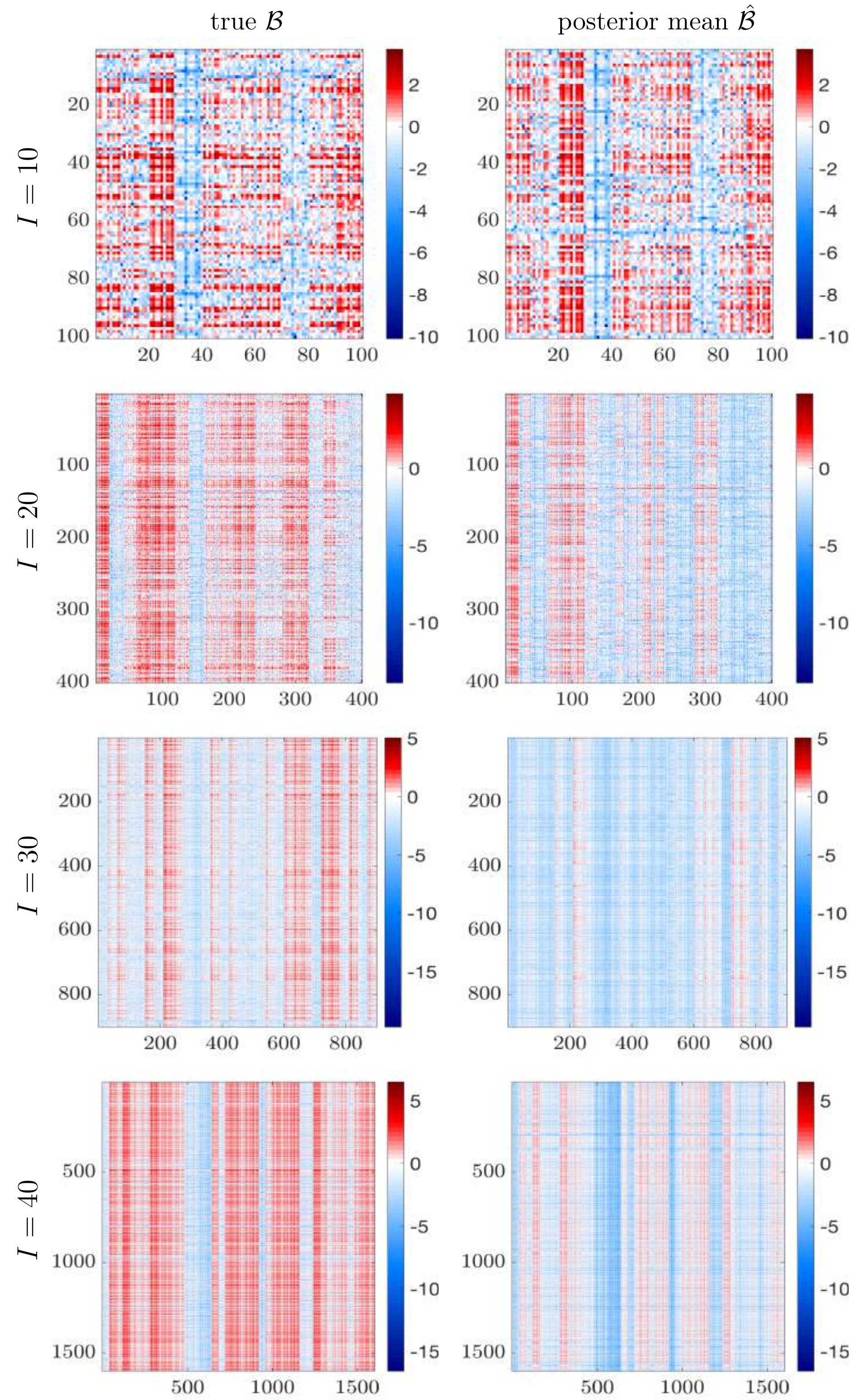

Figure 5: Logarithm of the absolute value of the coefficient tensors (in matricized form): true $\mathcal{B}$ (*left*) and posterior mean estimate $\hat{\mathcal{B}}$ (*right*), for four experiments with different size I (*in row*).

S.8.1 Experiment: I=10

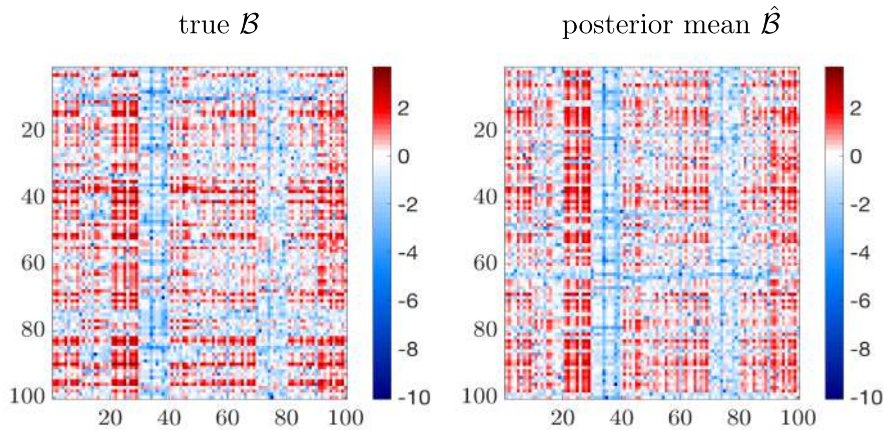

Figure 6: Experiment $I = 10$. Logarithm of the absolute value of the coefficient tensors (in matricized form): true $\mathcal{B}$ (*left*) and posterior mean estimate $\hat{\mathcal{B}}$ (*right*).

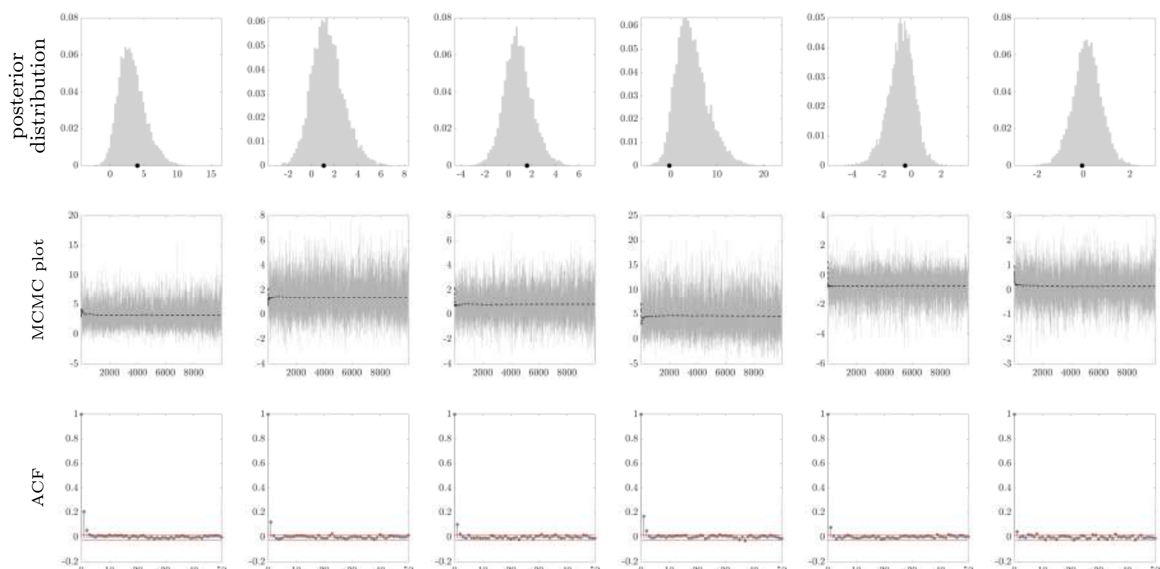

Figure 7: Experiment $I = 10$. Posterior distribution (*first row*, the black dot is the true value), MCMC plot (*second row*, dashed line represents the progressive mean) and autocorrelation function (*third row*) for some randomly chosen cells of the estimated coefficient tensor $\hat{\mathcal{B}}$ (in each *column*).

S.8.2 Experiment: I=20

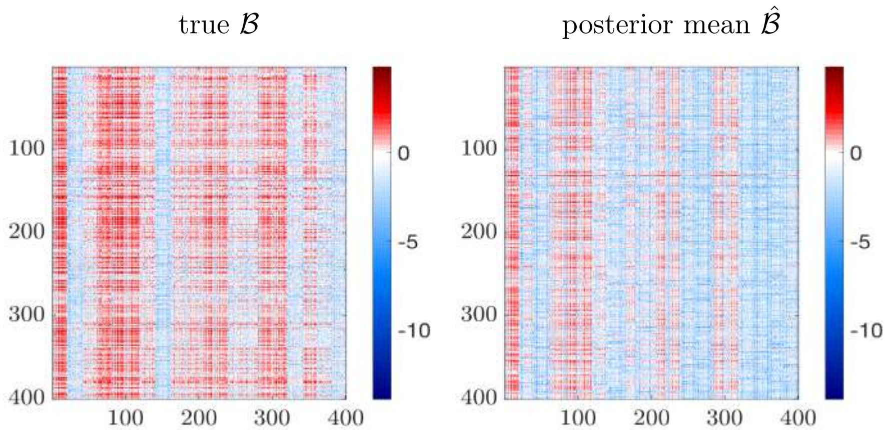

Figure 8: Experiment $I = 20$. Logarithm of the absolute value of the coefficient tensors (in matricized form): true $\mathcal{B}$ (*left*) and posterior mean estimate $\hat{\mathcal{B}}$ (*right*).

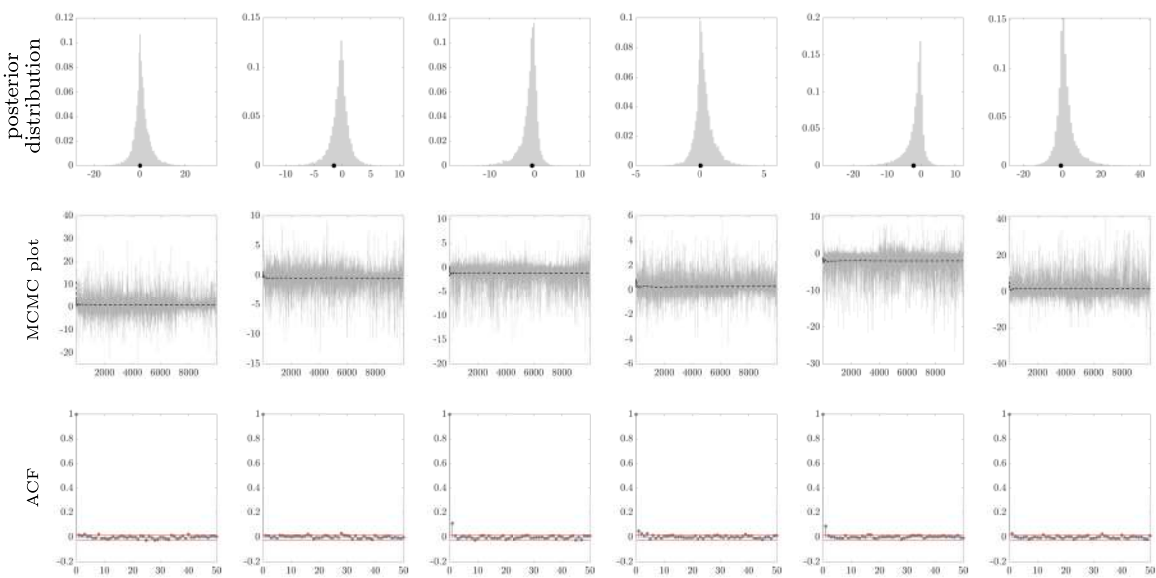

Figure 9: Experiment $I = 20$. Posterior distribution (*first row*, the black dot is the true value), MCMC plot (*second row*, dashed line represents the progressive mean) and autocorrelation function (*third row*) for some randomly chosen cells of the estimated coefficient tensor $\hat{\mathcal{B}}$ (in each *column*).

S.8.3 Experiment: I=30

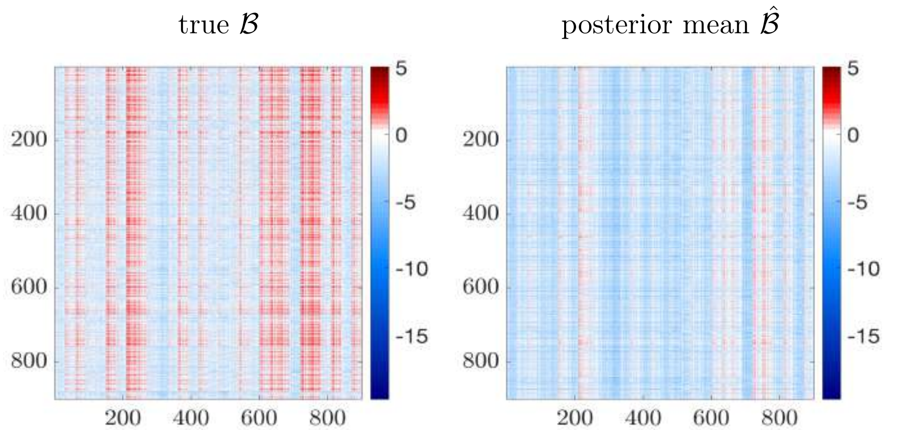

Figure 10: Experiment $I = 30$. Logarithm of the absolute value of the coefficient tensors (in matricized form): true $\mathcal{B}$ (*left*) and posterior mean estimate $\hat{\mathcal{B}}$ (*right*).

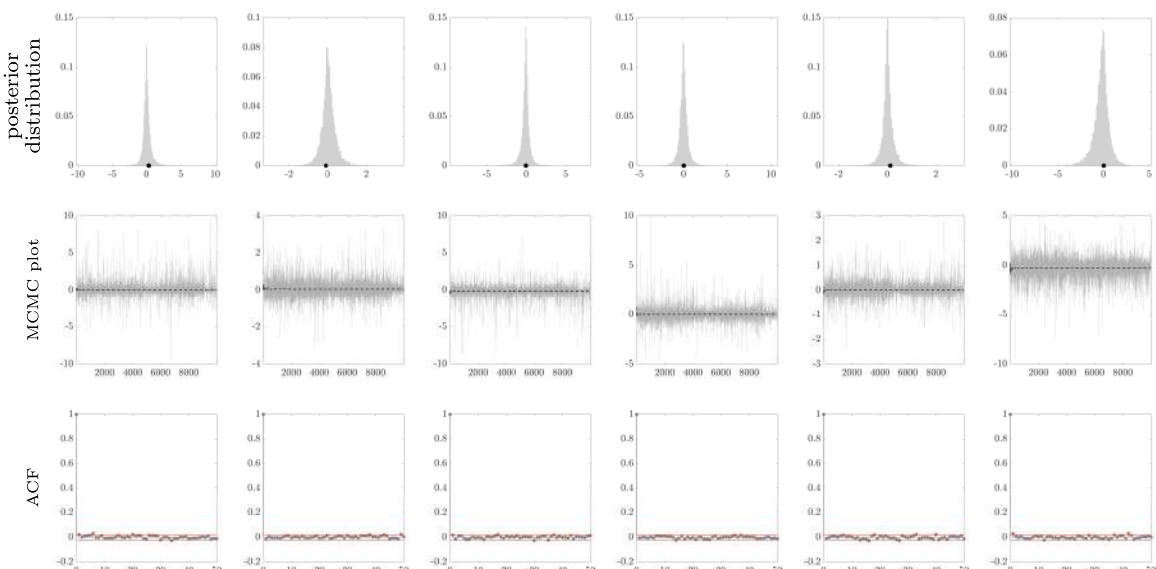

Figure 11: Experiment $I = 30$. Posterior distribution (*first row*, the black dot is the true value), MCMC plot (*second row*, dashed line represents the progressive mean) and autocorrelation function (*third row*) for some randomly chosen cells of the estimated coefficient tensor $\hat{\mathcal{B}}$ (in each *column*).

S.8.4 Experiment: I=40

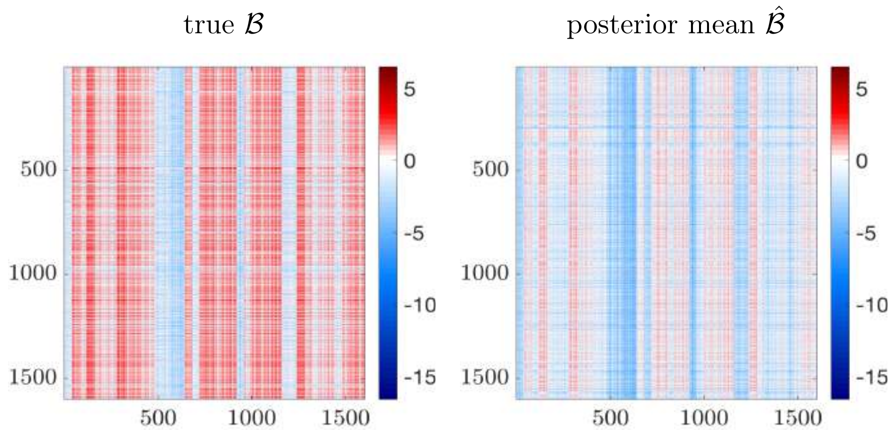

Figure 12: Experiment $I = 40$. Logarithm of the absolute value of the coefficient tensors (in matricized form): true $\mathcal{B}$ (*left*) and posterior mean estimate $\hat{\mathcal{B}}$ (*right*).

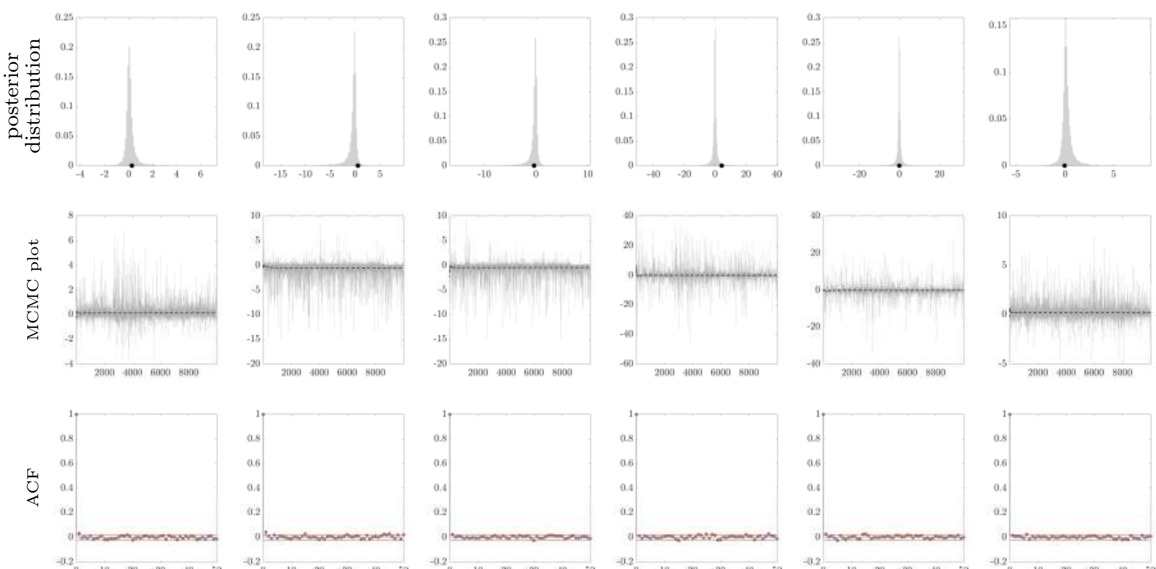

Figure 13: Experiment $I = 40$. Posterior distribution (*first row*, the black dot is the true value), MCMC plot (*second row*, dashed line represents the progressive mean) and autocorrelation function (*third row*) for some randomly chosen cells of the estimated coefficient tensor $\hat{\mathcal{B}}$ (in each *column*).

S.8.5 Comparison with competing models

In this section we compare the performance of the ART model (ART) proposed in Section 1.2 of the main paper against several alternatives models differing in terms of the shrinkage prior. Among the VAR models, we consider (i) a VAR with Dirichlet-Laplace prior (VAR-DL); (ii) a VAR with Horseshoe prior (VAR-HS); and (iii) a VAR with Normal-Gamma prior (VAR-NG). Among the univariate models, we consider: (i) an ARX with Elastic Net prior (ARX-EN); (ii) an ARX with Fused Lasso prior (ARX-FL); and (iii) an ARX with Normal-Gamma prior (ARX-NG).

We considered several synthetic datasets in this simulation setting. All of them have been generated as follows. We have simulated a $\mathcal{Y}_1$ from an order-3 tensor Normal distribution and we have specified the covariance matrices Σ_1, Σ_2, and Σ_3 in order to have high cross-correlations. Then, we have specified the entries of the coefficient tensor $\mathcal{B}$ without referring to the PARAFAC(R) decomposition, by fixing each entry $\mathcal{B}$ to a given value. Finally, we have generated the tensor $\mathcal{Y}_t$ by drawing from the ART(1) model.

For each synthetic dataset, we have used different sizes of the simulated data. In particular, we have fixed $I = J$, $K = 2$, and $T = 100$, then varied the size I across datasets.

The coefficient tensor $\mathcal{B}$ has been specified according to various instances of *partial heterogeneity*. With use this term to denote the case in which the entries of the coefficient tensor can be divided into groups such that coefficients have similar values within groups, but differ across groups, for example when the coefficient values have heterogeneity across covariates, with partial pooling within blocks of nodes for each covariate. In particular, we considered the following scenarios:

- scenario "col", where the first $I/2$ rows are set to 0.1, while the remaining $I/2$ rows are set to -0.1. In formulas

$$\mathcal{B}(1:I/2,:,:) = 0.1, \qquad \mathcal{B}(I/2+1:I,:,:) = -0.1$$

- scenario "row", where the first $J/2$ columns are set to 0.1, while the remaining $J/2$ columns are set to -0.1. In formulas

$$\mathcal{B}(:,1:J/2,:,:) = 0.1, \qquad \mathcal{B}(:,J/2+1:J,:,:) = -0.1$$

- scenario "block", where the first $I/2$ rows and columns are set to 0.1, while the last $I/2$ rows and columns are set to -0.1. All remaining entries are set to 0 in both regimes. In formulas

$$\mathcal{B}(1:I/2, 1:I/2, :) = 0.1, \qquad \mathcal{B}(I/2+1:I, I/2+1:I, :) = -0.1$$

We estimate the following models: (i) the ART model (ART) proposed in Section 1.2, with $R = 2$; (ii) a VAR with Dirichlet-Laplace prior (VAR-DL); (iii) a VAR with Horseshoe prior (VAR-HS); (iv) a VAR with Normal-Gamma prior (VAR-NG); (v) an ARX with Elastic Net prior (ARX-EN); (vi) an ARX with Fused Lasso prior (ARX-FL); and (vii) an ARX with Normal-Gamma prior (ARX-NG). The performance of each model is assessed and compared using the Deviance Information Criterion (DIC) of Spiegelhalter et al. (2002):

$$\mathrm{DIC} = -4\mathbb{E}_{\theta|\mathcal{Y}}\big[\log(p(\mathcal{Y}|\theta))\big] + 2\log(p(\mathcal{Y}|\tilde{\theta})),$$

where $\tilde{\theta}$ is an estimate of the parameters θ based on $\mathcal{Y}$. Since our framework involves hierarchical priors, we adopt the modifications of the DIC introduced by Celeux et al. (2006) and apply their observed DIC_3, which is the most reliable criterion:

$$\mathrm{DIC}_3 = -4\mathbb{E}_{\theta|\mathcal{Y}}\big[\log(p(\mathcal{Y}|\theta))\big] + 2\log(\hat{p}(\mathcal{Y})),$$

where $\hat{p}(\mathcal{Y}) \approx \mathbb{E}_{\theta|\mathcal{Y}}[p(\mathcal{Y}|\theta)]$ is an estimate of the density $p(\mathcal{Y}|\theta)$. Also, we consider the observed DIC_1 and DIC_2, defined as

$$\begin{aligned}
\mathrm{DIC}_1 &= -4\mathbb{E}_{\theta|\mathcal{Y}}\big[\log(p(\mathcal{Y}|\theta))\big] + 2\log(p(\mathcal{Y}|\mathbb{E}_{\theta|\mathcal{Y}}[\theta])) \\
\mathrm{DIC}_2 &= -4\mathbb{E}_{\theta|\mathcal{Y}}\big[\log(p(\mathcal{Y}|\theta))\big] + 2\log(p(\mathcal{Y}|\widehat{\theta}(\mathcal{Y}))),
\end{aligned}$$

where $\widehat{\theta}(\mathcal{Y}) = \arg\max_\theta p(\theta|\mathcal{Y})$. The lowest value of the DIC is associated to the best performing model. Tables 1 to 4 report the average DICs across $N = 4$ independent runs of the MCMC algorithms. They show that in all synthetic datasets the ART model outperforms the alternatives according to almost all DIC criteria, and it is always better when considering the average of the three DICs.

	Scenario "col"			Scenario "row"			Scenario "block"		
	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3
ART	576.706	562.462	563.991	447.997	428.820	429.878	45.375	26.706	28.354
VAR-DL	13732.777	13732.777	13191.025	13047.782	13047.782	12393.167	14644.784	14644.784	13775.223
VAR-HS	9320.585	9320.585	8728.634	9624.684	9624.684	8995.342	9742.699	9742.699	9033.025
VAR-NG	25912.969	25912.969	25551.393	26019.359	26019.359	25668.336	30067.059	30067.059	29536.112
ARX-FL	14702.737	16882.066	16731.699	14734.880	16897.955	16764.594	15089.136	17252.961	17121.347
ARX-EN	8518.901	8448.587	8507.715	8829.868	8773.106	8813.463	8875.289	8840.409	8860.338
ARX-NG	1088.602	8650.369	8775.561	2240.442	8988.962	9045.484	2691.705	9000.719	9043.601

Table 1: DIC for all datasets with $I = J = 3$, $K = 2$, all models. For each criterion, the best performing model is shaded in gray.

	Scenario "col"			Scenario "row"			Scenario "block"		
	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3
ART	-4130.588	-4195.275	-4175.689	-1440.704	-1493.073	-1463.639	-2236.844	-2311.603	-2293.562
VAR-DL	193153.847	193153.847	114374.380	103119.915	103119.915	69868.471	155665.852	155665.852	99708.231
VAR-HS	40851.416	40851.416	25001.917	39413.324	39413.324	24289.906	40680.997	40680.997	24950.047
VAR-NG	46447.674	46447.674	30469.139	86828.164	86828.164	71544.706	61973.283	61973.283	45086.537
ARX-FL	59331.359	59309.024	62652.875	56991.926	57209.369	60176.461	57245.086	57224.722	60506.492
ARX-EN	24492.253	24313.630	24518.891	23763.461	23597.865	23770.431	23978.278	23791.146	24015.639
ARX-NG	3101.543	24847.157	25619.932	4775.989	24269.654	24715.470	3791.682	24029.922	25093.752

Table 2: DIC for all datasets with $I = J = 5$, $K = 2$, all models. For each criterion, the best performing model is shaded in gray.

	Scenario "col"			Scenario "row"			Scenario "block"		
	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3
ART	-10774.609	-10916.236	-10879.313	-14013.939	-14138.553	-14102.565	-11675.367	-11791.712	-11741.614
VAR-DL	264156.976	264156.976	137049.407	264436.246	264436.246	136922.979	264445.144	264445.144	137026.286
VAR-HS	116535.928	116535.928	62925.810	118131.221	118131.221	63704.893	115765.060	115765.060	62515.267
VAR-NG	118977.014	118977.014	64556.261	119301.103	119301.103	64507.923	117566.926	117566.926	63767.642
ARX-FL	179641.897	179480.821	185338.733	183019.728	182870.040	188851.264	181368.732	181130.233	187101.038
ARX-EN	60075.777	59157.424	60829.314	62188.741	61228.386	62905.239	61657.391	60894.869	62560.377
ARX-NG	22907.275	49949.312	62568.541	24346.027	54898.976	68940.280	18628.992	54167.939	65159.375

Table 3: DIC for all datasets with $I = J = 8$, $K = 2$, all models. For each criterion, the best performing model is shaded in gray.

	Scenario "col"			Scenario "row"			Scenario "block"		
	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3
ART	-18717.110	-18906.844	-18851.326	-23230.938	-23388.626	-23323.538	-25188.265	-25359.805	-25325.457
VAR-DL	264428.872	264428.872	136773.555	264428.872	264428.872	136773.555	264428.872	264428.872	136773.555
VAR-HS	180685.575	180685.575	95059.618	186572.219	186572.219	97945.097	184316.464	184316.464	96843.670
VAR-NG	183966.476	183966.476	97127.248	188222.521	188222.521	98995.114	187376.441	187376.441	98876.450
ARX-FL	302008.728	301636.906	310983.438	304527.696	304119.870	313481.373	298638.207	298584.145	307400.599
ARX-EN	93203.598	91812.811	95564.926	101038.793	99992.610	103427.286	98411.973	97457.887	101038.399
ARX-NG	10534.692	63325.926	96507.632	17667.199	58921.913	89898.942	11233.856	58966.954	92380.492

Table 4: DIC for all datasets with $I = J = 10$, $K = 2$, all models. For each criterion, the best performing model is shaded in gray.

	Scenario "col"			Scenario "row"			Scenario "block"		
	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3	DIC_1	DIC_2	DIC_3
ART	-42124.402	-42502.912	-42239.065	-46822.431	-47121.484	-46999.071	-39833.172	-40066.877	-40010.286
VAR-DL	264428.872	264428.872	136773.555	264428.872	264428.872	136773.555	264428.872	264428.872	136773.555
VAR-HS	127069.799	127069.799	66062.957	255063.736	255063.736	132443.465	251330.055	251330.055	130620.693
VAR-NG	258589.814	258589.814	134557.032	258630.605	258630.605	134353.769	190124.950	190124.950	98831.861
ARX-FL	455541.820	454865.670	469672.054	462143.716	461487.288	476560.986	456205.086	455589.243	470448.549
ARX-EN	139515.266	136489.369	144225.795	149218.505	148158.164	155012.392	140842.674	139422.444	145896.572
ARX-NG	30576.543	74265.572	113990.145	21490.735	75676.871	122422.809	36616.317	76137.700	118662.140

Table 5: DIC for all datasets with $I = J = 12$, $K = 2$, all models. For each criterion, the best performing model is shaded in gray.

S.9 Data Description

As put forward by Schweitzer et al. (2009), the analysis of economic networks is one of the most recent and complex challenges that the econometric community is facing nowadays. We contribute to the econometric literature about complex networks by applying the proposed methodology to the study jointly the dynamics of international trade and credit networks. The international trade and financial networks have been previously studied by several authors (e.g., see Anundsen et al., 2016; Eaton and Kortum, 2002; Fagiolo et al., 2009; Fieler, 2011; Hidalgo and Hausmann, 2009; Kharrazi et al., 2017; Meyfroidt et al., 2010; Squartini et al., 2011; Zhu et al., 2014), who investigated its topological properties and identified its main communities. To the best of our knowledge, this is the first attempt to model the dynamics of two networks jointly.

The bilateral trade data come from the United Nations COMTRADE database[3], whereas the data on bilateral outstanding capital come from the Bank of International Settlements database[4], both are publicly available resources. For each couple (i, j) of countries, the international trade data from COMTRADE report total exports from country i to country j occurred during year t, while the BIS dataset gives the total amount of claims (i.e., credit) of country i vis-á-vis country j in year t. We use a subset of the COMTRADE database. Our sample of yearly observations for 10 countries ($I_1 = I_2 = I = 10$) runs from 2003 to 2016. In order to remove potential non-linearities in the data, we take the logarithm all variables of interest. We thus consider the international trade and financial network in each period as one observation from a real-valued tensor-valued stochastic process. To sum up, our dataset consists in a 3-order tensor-valued time series of length $T = 13$. At each time t, the 3-order tensor $\mathcal{Y}_t$ has dimension (I_1, I_2, I_3), with $I_1 = I_2 = I = 10$ and $I_3 = 2$, and it represents a 2-layer node-aligned network (or multiplex) with 10 vertices (countries), where each edge is given by a bilateral trade flow or financial stock. The entry $(i, j, 1, t)$ of $\mathcal{Y}_t$ reports the total exports of country i vis-à-vis country j, in year t, whereas entry $(i, j, 2, t)$ contains the total outstanding credit from country i towards country j, in year t. The series $(\mathcal{Y}_t)_t$, $t = 1, \ldots, T$, has been standardized (over the temporal dimension).

S.10 Additional results for the empirical application

S.10.1 Estimation results

Fig. 16 shows the estimated covariance matrices. In all cases, the highest values correspond to individual variances, while the estimated covariances are lower in magnitude and heterogeneous. We also find evidence of heterogeneity in the dependence structure, since Σ_1, which captures the covariance between rows (i.e., exporting and creditor countries), differs from Σ_2, which describes the covariance between columns (i.e., importing and debtor countries). With few exceptions, estimated covariances are positive.

[3] https://comtrade.un.org

[4] http://stats.bis.org/statx/toc/LBS.html

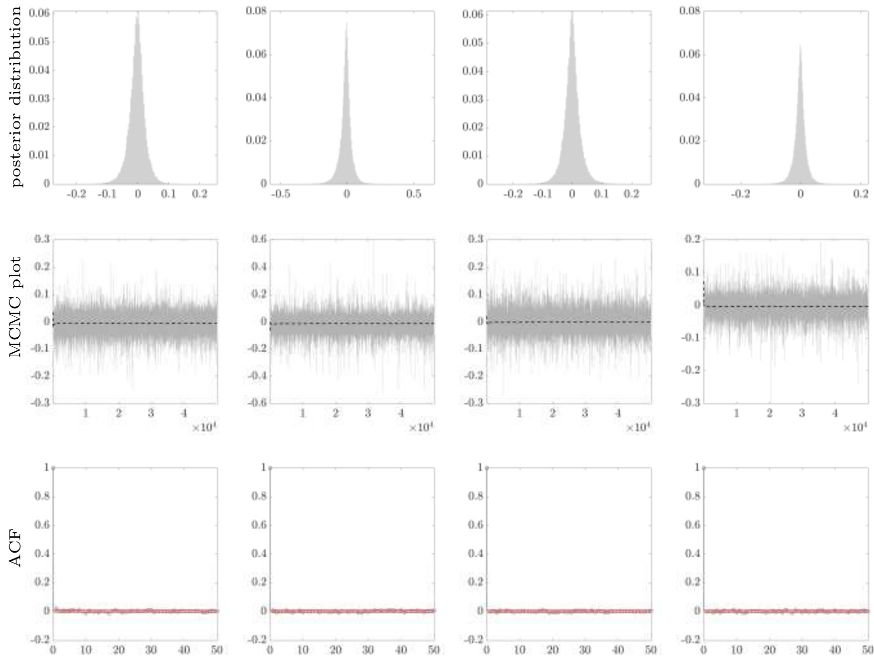

Figure 14: Posterior distribution (*first row*), MCMC plot (*second row*, dashed line represents the progressive mean) and autocorrelation function (*third row*) for four randomly chosen cells of the estimated coefficient tensor $\hat{\mathcal{B}}$ (in each *column*).

To assess the convergence of the MCMC algorithm, we have performed a convergence diagnostics analysis based on the `coda` functions of the LeSage' Econometrics toolbox[5] (LeSage, 1999). Specifically, we rely on the diagnostic criteria of the Geweke (1992) and Raftery and Lewis (1995).

Raftery and Lewis' approach allows to determine how long to monitor the chain in order to achieve a pre-specified level of accuracy of the posterior summaries. The default values require that, for nominal reporting based on a 95% interval using the 0.025 and 0.975 quantile points, the actual posterior values should result lie between 0.95 and 0.96. The results of this procedure consists in the thinning factor, the burn-in period, and the total number of draws (N) needed to achieve the desired accuracy of the sampler. Also, the fourth column reports the number of draws that would be needed if the draws represented an i.i.d. chain ($Nmin$). Finally, the I-statistic, which is given by the ratio of the third to the fourth column (i.e., $N/Nmin$), provides evidence of convergence problems if its values exceeds 5.

[5]See also `https://www.spatial-econometrics.com`.

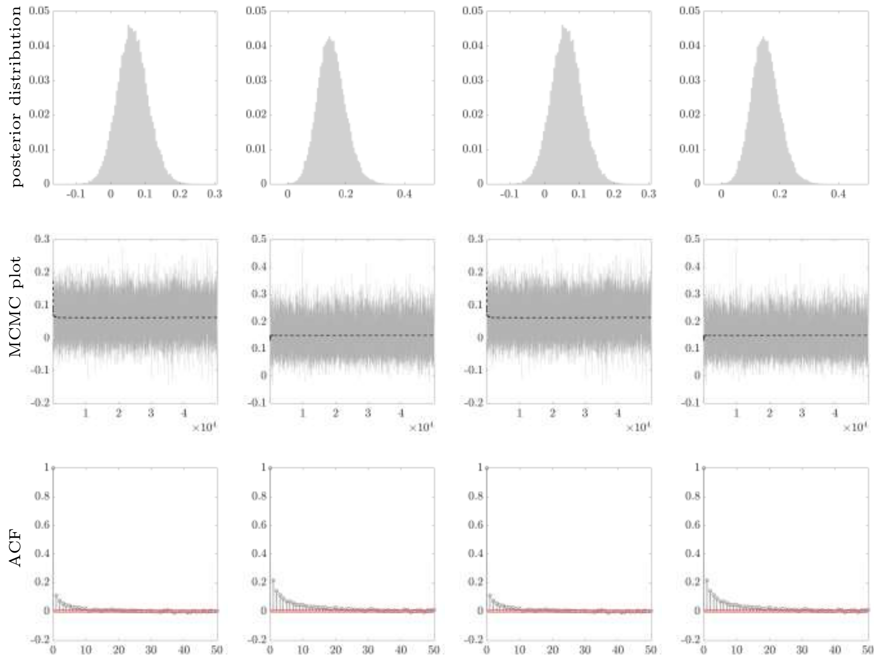

Figure 15: Posterior distribution (*first row*), MCMC plot (*second row*, dashed line represents the progressive mean) and autocorrelation function (*third row*) for two randomly chosen cells of the estimated covariance matrix $\hat{\Sigma}_1$ (*first and second column*) and $\hat{\Sigma}_2$ (*third and fourth column*).

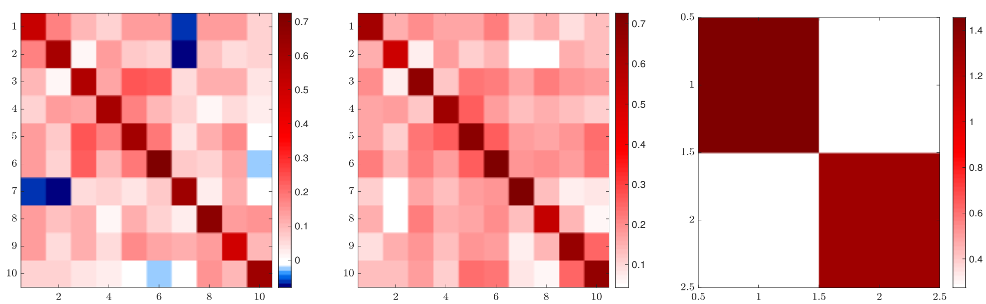

Figure 16: Estimated covariance matrices: $\hat{\Sigma}_1$ (*left*), $\hat{\Sigma}_2$ (*center*), $\hat{\Sigma}_3$ (*right*).

The Geweke diagnostics consists in the estimates of the numerical standard errors (NSE) and relative numerical efficiency (RNE), based on the assumption that the draws come from an i.i.d. process (first column), as well as on a 4%, 8%, and 15% tapering (or truncation) of the periodgram window used to approximate the spectral density of the parameter of interest (second to fourth column). This second set of columns take into account the autocorrelation among the MCMC draws, thus one should rely on them in case of disagreement with the i.i.d. estimates. To interpret the results, notice that the RNE provides an estimate of the number of MCMC draws that would be required to produce the same numerical accuracy as if the draws had been made from an i.i.d. sample drawn from the posterior distribution. Therefore, values of the RNE close to unity are indicative of the i.i.d. nature of our sample.

Moreover, Geweke proposed a test for assessing whether the chain for a given parameter has converged that is, it has reached the equilibrium. Specifically, he designed a Z-test for hypothesis of equality of the means computed using the first 20% and the last 50% of the draws from the chain. The resulting p-value greater than α implies the non-rejection of the null (of convergence) at the α confidence level.

We compute these measures, together with the autocorrelation function at lags $1, 5, 10, 50$, for some randomly chosen entries of the coefficient tensor, $\hat{\mathcal{B}}$, and the covariance matrices, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$. The results reported in Tables 6 to 14 provide evidence of convergence according to all criteria each of the parameters.

Autocorrelation function				
	lag 1	lag 5	lag 10	lag 50
$\hat{\mathcal{B}}_{9,3,1,169}$	0.041	0.020	-0.015	-0.001
$\hat{\mathcal{B}}_{2,8,2,163}$	-0.010	0.011	-0.001	0.010
$\hat{\mathcal{B}}_{10,9,2,192}$	-0.024	-0.015	-0.011	-0.012
$\hat{\mathcal{B}}_{5,5,2,28}$	-0.005	-0.007	0.011	-0.020
$\hat{\mathcal{B}}_{2,8,2,30}$	0.023	0.007	0.025	0.010
$\hat{\mathcal{B}}_{9,4,1,102}$	0.056	-0.008	-0.017	-0.022
$\hat{\mathcal{B}}_{2,8,2,51}$	0.023	-0.025	0.020	0.014
$\hat{\mathcal{B}}_{4,7,2,140}$	-0.022	-0.012	0.008	-0.004
$\hat{\mathcal{B}}_{1,5,2,49}$	0.018	0.001	0.011	-0.022
$\hat{\mathcal{B}}_{7,4,1,179}$	0.005	0.013	0.007	-0.004
$\hat{\mathcal{B}}_{6,2,1,69}$	0.008	-0.010	-0.011	-0.025
$\hat{\mathcal{B}}_{1,6,1,151}$	0.026	0.018	-0.008	-0.028
$\hat{\mathcal{B}}_{1,1,2,52}$	-0.010	0.011	0.016	-0.015
$\hat{\mathcal{B}}_{3,6,2,45}$	-0.016	-0.016	0.024	-0.037
$\hat{\mathcal{B}}_{1,2,1,118}$	0.024	0.009	-0.021	0.029
$\hat{\mathcal{B}}_{9,9,1,110}$	0.006	-0.003	0.001	-0.010
$\hat{\mathcal{B}}_{1,8,2,186}$	0.013	0.046	-0.007	0.005
$\hat{\mathcal{B}}_{4,1,1,52}$	0.039	0.018	-0.010	0.015
$\hat{\mathcal{B}}_{10,10,2,70}$	0.002	0.016	0.009	0.007
$\hat{\mathcal{B}}_{2,8,2,192}$	-0.013	-0.016	-0.028	-0.006

Raftery and Lewis diagnostics					
	thin	burn	total (N)	Nmin	I-stat
$\hat{\mathcal{B}}_{9,3,1,169}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{2,8,2,163}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{10,9,2,192}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{5,5,2,28}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{2,8,2,30}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{9,4,1,102}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{2,8,2,51}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{4,7,2,140}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{1,5,2,49}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{7,4,1,179}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{6,2,1,69}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{1,6,1,151}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{1,1,2,52}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{3,6,2,45}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{1,2,1,118}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{9,9,1,110}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{1,8,2,186}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{4,1,1,52}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{10,10,2,70}$	1.000	2.000	984.000	937.000	1.050
$\hat{\mathcal{B}}_{2,8,2,192}$	1.000	2.000	984.000	937.000	1.050

Table 6: Convergence diagnostics for randomly selected entries of $\hat{\mathcal{B}}$: autocorrelation function (left), Raftery and Lewis convergence diagnostics (right).

Geweke diagnostics								
	NSE iid	RNE iid	NSE 4%	RNE 4%	NSE 8%	RNE 8%	NSE 15%	RNE 15%
$\hat{\mathcal{B}}_{9,3,1,169}$	0.000	1.000	0.000	1.001	0.000	1.011	0.000	1.221
$\hat{\mathcal{B}}_{2,8,2,163}$	0.001	1.000	0.001	0.828	0.001	0.811	0.001	0.732
$\hat{\mathcal{B}}_{10,9,2,192}$	0.000	1.000	0.000	1.869	0.000	2.648	0.000	2.720
$\hat{\mathcal{B}}_{5,5,2,28}$	0.001	1.000	0.001	1.217	0.001	1.442	0.000	2.151
$\hat{\mathcal{B}}_{2,8,2,30}$	0.001	1.000	0.001	0.554	0.001	0.420	0.001	0.380
$\hat{\mathcal{B}}_{9,4,1,102}$	0.000	1.000	0.000	0.833	0.000	0.874	0.000	0.800
$\hat{\mathcal{B}}_{2,8,2,51}$	0.001	1.000	0.001	1.048	0.001	0.935	0.001	0.834
$\hat{\mathcal{B}}_{4,7,2,140}$	0.000	1.000	0.000	0.913	0.000	0.870	0.000	1.224
$\hat{\mathcal{B}}_{1,5,2,49}$	0.001	1.000	0.001	0.871	0.001	0.809	0.001	0.836
$\hat{\mathcal{B}}_{7,4,1,179}$	0.000	1.000	0.000	0.855	0.000	0.892	0.000	0.923
$\hat{\mathcal{B}}_{6,2,1,69}$	0.000	1.000	0.000	1.058	0.000	1.173	0.000	1.490
$\hat{\mathcal{B}}_{1,6,1,151}$	0.000	1.000	0.000	0.927	0.000	1.371	0.000	2.041
$\hat{\mathcal{B}}_{1,1,2,52}$	0.001	1.000	0.001	1.095	0.001	1.072	0.001	1.004
$\hat{\mathcal{B}}_{3,6,2,45}$	0.000	1.000	0.000	1.131	0.000	1.161	0.000	1.367
$\hat{\mathcal{B}}_{1,2,1,118}$	0.001	1.000	0.001	0.956	0.001	0.901	0.001	0.941
$\hat{\mathcal{B}}_{9,9,1,110}$	0.000	1.000	0.000	1.095	0.000	1.182	0.000	1.377
$\hat{\mathcal{B}}_{1,8,2,186}$	0.001	1.000	0.001	0.830	0.001	0.989	0.001	1.209
$\hat{\mathcal{B}}_{4,1,1,52}$	0.000	1.000	0.000	1.046	0.000	1.382	0.000	2.434
$\hat{\mathcal{B}}_{10,10,2,70}$	0.000	1.000	0.000	1.377	0.000	1.603	0.000	2.014
$\hat{\mathcal{B}}_{2,8,2,192}$	0.001	1.000	0.001	1.768	0.000	2.359	0.000	2.553

Table 7: Geweke convergence diagnostics for randomly selected entries of $\hat{\mathcal{B}}$.

Geweke's test												
	i.i.d.			4% taper			8% taper			15% taper		
	mean	NSE	p-value	mean	NSE	p-value	mean	NSE	p-value	mean	NSE	p-value
$\hat{\mathcal{B}}_{9,3,1,169}$	0.003	0.000	0.643	0.003	0.000	0.669	0.003	0.000	0.673	0.003	0.000	0.657
$\hat{\mathcal{B}}_{2,8,2,163}$	0.007	0.001	0.246	0.006	0.001	0.234	0.006	0.001	0.212	0.006	0.001	0.180
$\hat{\mathcal{B}}_{10,9,2,192}$	-0.004	0.000	0.884	-0.004	0.000	0.867	-0.004	0.000	0.859	-0.004	0.000	0.833
$\hat{\mathcal{B}}_{5,5,2,28}$	-0.002	0.001	0.889	-0.002	0.001	0.885	-0.002	0.001	0.883	-0.002	0.001	0.889
$\hat{\mathcal{B}}_{2,8,2,30}$	-0.011	0.001	0.068	-0.010	0.001	0.109	-0.010	0.001	0.144	-0.010	0.001	0.180
$\hat{\mathcal{B}}_{9,4,1,102}$	0.004	0.000	0.154	0.004	0.000	0.204	0.004	0.000	0.212	0.004	0.000	0.127
$\hat{\mathcal{B}}_{2,8,2,51}$	0.002	0.001	0.063	0.002	0.001	0.097	0.002	0.001	0.073	0.002	0.001	0.039
$\hat{\mathcal{B}}_{4,7,2,140}$	0.004	0.000	0.532	0.004	0.000	0.567	0.004	0.000	0.578	0.004	0.000	0.557
$\hat{\mathcal{B}}_{1,5,2,49}$	-0.006	0.001	0.543	-0.006	0.001	0.515	-0.006	0.001	0.538	-0.006	0.001	0.560
$\hat{\mathcal{B}}_{7,4,1,179}$	-0.000	0.000	0.996	-0.000	0.000	0.996	-0.000	0.000	0.996	-0.000	0.000	0.996
$\hat{\mathcal{B}}_{6,2,1,69}$	0.000	0.000	0.217	0.000	0.000	0.287	0.000	0.000	0.288	0.000	0.000	0.292
$\hat{\mathcal{B}}_{1,6,1,151}$	-0.006	0.000	0.571	-0.006	0.000	0.616	-0.006	0.000	0.646	-0.006	0.000	0.648
$\hat{\mathcal{B}}_{1,1,2,52}$	-0.003	0.001	0.630	-0.003	0.001	0.619	-0.003	0.001	0.620	-0.003	0.001	0.637
$\hat{\mathcal{B}}_{3,6,2,45}$	0.000	0.001	0.358	-0.000	0.000	0.255	0.000	0.000	0.269	0.000	0.000	0.276
$\hat{\mathcal{B}}_{1,2,1,118}$	-0.003	0.001	0.196	-0.003	0.001	0.219	-0.003	0.001	0.209	-0.003	0.001	0.189
$\hat{\mathcal{B}}_{9,9,1,110}$	-0.000	0.000	0.089	-0.000	0.000	0.068	-0.000	0.000	0.069	-0.000	0.000	0.036
$\hat{\mathcal{B}}_{1,8,2,186}$	-0.005	0.001	0.380	-0.005	0.001	0.410	-0.005	0.001	0.430	-0.005	0.001	0.461
$\hat{\mathcal{B}}_{4,1,1,52}$	-0.002	0.000	0.980	-0.002	0.000	0.979	-0.002	0.000	0.980	-0.002	0.000	0.978
$\hat{\mathcal{B}}_{10,10,2,70}$	-0.002	0.000	0.199	-0.001	0.000	0.226	-0.001	0.000	0.195	-0.001	0.000	0.133
$\hat{\mathcal{B}}_{2,8,2,192}$	0.009	0.001	0.697	0.010	0.001	0.628	0.009	0.001	0.607	0.009	0.001	0.589

Table 8: Geweke's test for randomly selected entries of $\hat{\mathcal{B}}$.

Autocorrelation function				
	lag 1	lag 5	lag 10	lag 50
$\hat{\Sigma}_{1,7,9}$	0.187	0.069	0.039	-0.020
$\hat{\Sigma}_{1,7,3}$	0.148	0.065	0.027	-0.034
$\hat{\Sigma}_{1,5,6}$	0.229	0.065	0.005	-0.036
$\hat{\Sigma}_{1,8,6}$	0.105	0.002	0.021	0.022
$\hat{\Sigma}_{1,1,9}$	0.207	0.083	0.047	-0.008
$\hat{\Sigma}_{1,5,2}$	0.210	0.044	0.019	0.019
$\hat{\Sigma}_{1,6,9}$	0.182	0.078	0.014	-0.028

Raftery and Lewis diagnostics					
	thin	burn	total (N)	Nmin	I-stat
$\hat{\Sigma}_{1,7,9}$	1.000	3.000	1035.000	937.000	1.105
$\hat{\Sigma}_{1,7,3}$	1.000	3.000	1035.000	937.000	1.105
$\hat{\Sigma}_{1,5,6}$	1.000	3.000	1035.000	937.000	1.105
$\hat{\Sigma}_{1,8,6}$	1.000	3.000	1035.000	937.000	1.105
$\hat{\Sigma}_{1,1,9}$	1.000	3.000	1035.000	937.000	1.105
$\hat{\Sigma}_{1,5,2}$	1.000	3.000	1035.000	937.000	1.105
$\hat{\Sigma}_{1,6,9}$	1.000	3.000	1035.000	937.000	1.105

Table 9: Convergence diagnostics for randomly selected entries of $\hat{\Sigma}_1$: autocorrelation function (left), Raftery and Lewis convergence diagnostics (right).

Geweke diagnostics								
	NSE iid	RNE iid	NSE 4%	RNE 4%	NSE 8%	RNE 8%	NSE 15%	RNE 15%
$\hat{\Sigma}_{1,7,9}$	0.001	1.000	0.001	0.290	0.001	0.286	0.001	0.275
$\hat{\Sigma}_{1,7,3}$	0.001	1.000	0.001	0.345	0.001	0.364	0.001	0.455
$\hat{\Sigma}_{1,5,6}$	0.001	1.000	0.001	0.415	0.001	0.617	0.001	0.717
$\hat{\Sigma}_{1,8,6}$	0.001	1.000	0.001	0.622	0.001	0.590	0.001	0.567
$\hat{\Sigma}_{1,1,9}$	0.001	1.000	0.001	0.369	0.001	0.389	0.001	0.526
$\hat{\Sigma}_{1,5,2}$	0.001	1.000	0.001	0.309	0.001	0.367	0.001	0.540
$\hat{\Sigma}_{1,6,9}$	0.001	1.000	0.001	0.407	0.001	0.359	0.001	0.293

Table 10: Geweke convergence diagnostics for randomly selected entries of $\hat{\Sigma}_1$.

Geweke's test												
	i.i.d.			4% taper			8% taper			15% taper		
	mean	NSE	p-value	mean	NSE	p-value	mean	NSE	p-value	mean	NSE	p-value
$\hat{\Sigma}_{1,7,9}$	0.112	0.001	0.658	0.112	0.001	0.787	0.112	0.001	0.787	0.112	0.001	0.790
$\hat{\Sigma}_{1,7,3}$	0.057	0.001	0.191	0.057	0.001	0.420	0.057	0.001	0.423	0.057	0.001	0.404
$\hat{\Sigma}_{1,5,6}$	0.189	0.001	0.520	0.189	0.001	0.689	0.189	0.001	0.656	0.189	0.001	0.591
$\hat{\Sigma}_{1,8,6}$	0.067	0.001	0.687	0.067	0.001	0.777	0.067	0.001	0.770	0.067	0.001	0.754
$\hat{\Sigma}_{1,1,9}$	0.143	0.001	0.118	0.142	0.001	0.378	0.142	0.001	0.363	0.142	0.001	0.247
$\hat{\Sigma}_{1,5,2}$	0.073	0.001	0.519	0.072	0.001	0.699	0.073	0.001	0.700	0.073	0.001	0.699
$\hat{\Sigma}_{1,6,9}$	0.126	0.001	0.735	0.126	0.001	0.830	0.126	0.001	0.829	0.126	0.001	0.826

Table 11: Geweke's test for randomly selected entries of $\hat{\Sigma}_1$.

Autocorrelation function				
	lag 1	lag 5	lag 10	lag 50
$\hat{\Sigma}_{2,10,5}$	0.292	0.078	0.056	0.006
$\hat{\Sigma}_{2,9,8}$	0.210	0.036	0.035	0.013
$\hat{\Sigma}_{2,7,4}$	0.180	0.057	0.032	-0.015
$\hat{\Sigma}_{2,10,1}$	0.152	0.032	0.022	-0.002
$\hat{\Sigma}_{2,1,10}$	0.152	0.032	0.022	-0.002
$\hat{\Sigma}_{2,9,4}$	0.144	0.045	0.055	-0.004
$\hat{\Sigma}_{2,1,5}$	0.212	0.036	0.024	-0.006
$\hat{\Sigma}_{2,4,4}$	0.401	0.170	0.070	-0.041
$\hat{\Sigma}_{2,5,10}$	0.292	0.078	0.056	0.006
$\hat{\Sigma}_{2,9,5}$	0.163	0.023	0.028	-0.024
$\hat{\Sigma}_{2,5,3}$	0.249	0.077	0.028	-0.033

Raftery and Lewis diagnostics					
	thin	burn	total (N)	Nmin	I-stat
$\hat{\Sigma}_{2,10,5}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,9,8}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,7,4}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,10,1}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,1,10}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,9,4}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,1,5}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,4,4}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,5,10}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,9,5}$	1.000	3.000	1072.000	937.000	1.144
$\hat{\Sigma}_{2,5,3}$	1.000	3.000	1072.000	937.000	1.144

Table 12: Convergence diagnostics for randomly selected entries of $\hat{\Sigma}_2$: autocorrelation function (left), Raftery and Lewis convergence diagnostics (right).

Geweke diagnostics								
	NSE iid	RNE iid	NSE 4%	RNE 4%	NSE 8%	RNE 8%	NSE 15%	RNE 15%
$\hat{\Sigma}_{2,10,5}$	0.001	1.000	0.001	0.337	0.001	0.512	0.001	0.477
$\hat{\Sigma}_{2,9,8}$	0.001	1.000	0.001	0.329	0.001	0.335	0.001	0.429
$\hat{\Sigma}_{2,7,4}$	0.001	1.000	0.001	0.377	0.001	0.386	0.001	0.452
$\hat{\Sigma}_{2,10,1}$	0.001	1.000	0.001	0.501	0.001	0.484	0.001	0.404
$\hat{\Sigma}_{2,1,10}$	0.001	1.000	0.001	0.501	0.001	0.484	0.001	0.404
$\hat{\Sigma}_{2,9,4}$	0.001	1.000	0.001	0.585	0.001	0.683	0.001	0.860
$\hat{\Sigma}_{2,1,5}$	0.001	1.000	0.001	0.570	0.001	0.676	0.001	0.700
$\hat{\Sigma}_{2,4,4}$	0.001	1.000	0.002	0.281	0.002	0.417	0.002	0.502
$\hat{\Sigma}_{2,5,10}$	0.001	1.000	0.001	0.337	0.001	0.512	0.001	0.477
$\hat{\Sigma}_{2,9,5}$	0.001	1.000	0.001	0.712	0.001	0.939	0.001	0.975
$\hat{\Sigma}_{2,5,3}$	0.001	1.000	0.001	0.466	0.001	0.548	0.001	0.512

Table 13: Geweke convergence diagnostics for randomly selected entries of $\hat{\Sigma}_2$.

Geweke's test												
	i.i.d.			4% taper			8% taper			15% taper		
	mean	NSE	p-value	mean	NSE	p-value	mean	NSE	p-value	mean	NSE	p-value
$\hat{\Sigma}_{2,10,5}$	0.238	0.001	0.522	0.238	0.002	0.728	0.238	0.002	0.724	0.238	0.001	0.705
$\hat{\Sigma}_{2,9,8}$	0.141	0.001	0.278	0.141	0.001	0.525	0.141	0.001	0.539	0.141	0.001	0.462
$\hat{\Sigma}_{2,7,4}$	0.138	0.001	0.251	0.138	0.001	0.485	0.138	0.001	0.503	0.138	0.001	0.538
$\hat{\Sigma}_{2,10,1}$	0.133	0.001	0.214	0.133	0.001	0.390	0.133	0.001	0.377	0.132	0.001	0.314
$\hat{\Sigma}_{2,1,10}$	0.133	0.001	0.214	0.133	0.001	0.390	0.133	0.001	0.377	0.132	0.001	0.314
$\hat{\Sigma}_{2,9,4}$	0.131	0.001	0.396	0.131	0.001	0.554	0.131	0.001	0.590	0.131	0.001	0.607
$\hat{\Sigma}_{2,1,5}$	0.171	0.001	0.564	0.171	0.001	0.677	0.170	0.001	0.643	0.170	0.001	0.611
$\hat{\Sigma}_{2,4,4}$	0.637	0.001	0.714	0.637	0.003	0.875	0.637	0.003	0.879	0.637	0.002	0.872
$\hat{\Sigma}_{2,5,10}$	0.238	0.001	0.522	0.238	0.002	0.728	0.238	0.002	0.724	0.238	0.001	0.705
$\hat{\Sigma}_{2,9,5}$	0.190	0.001	0.436	0.190	0.001	0.614	0.190	0.001	0.628	0.190	0.001	0.631
$\hat{\Sigma}_{2,5,3}$	0.228	0.001	0.798	0.228	0.001	0.877	0.228	0.001	0.872	0.228	0.001	0.859

Table 14: Geweke's test for randomly selected entries of $\hat{\Sigma}_2$.

S.10.2 Impulse response analysis

Fig. 17 shows the block Cholesky IRF at horizon $h = 1, 2$, resulting from a negative 1% shock to GB's outstanding debt[6]. The main findings follow.

[6] Again, the shock is allocated across countries to reflect country-specific shares of the last period in the sample.

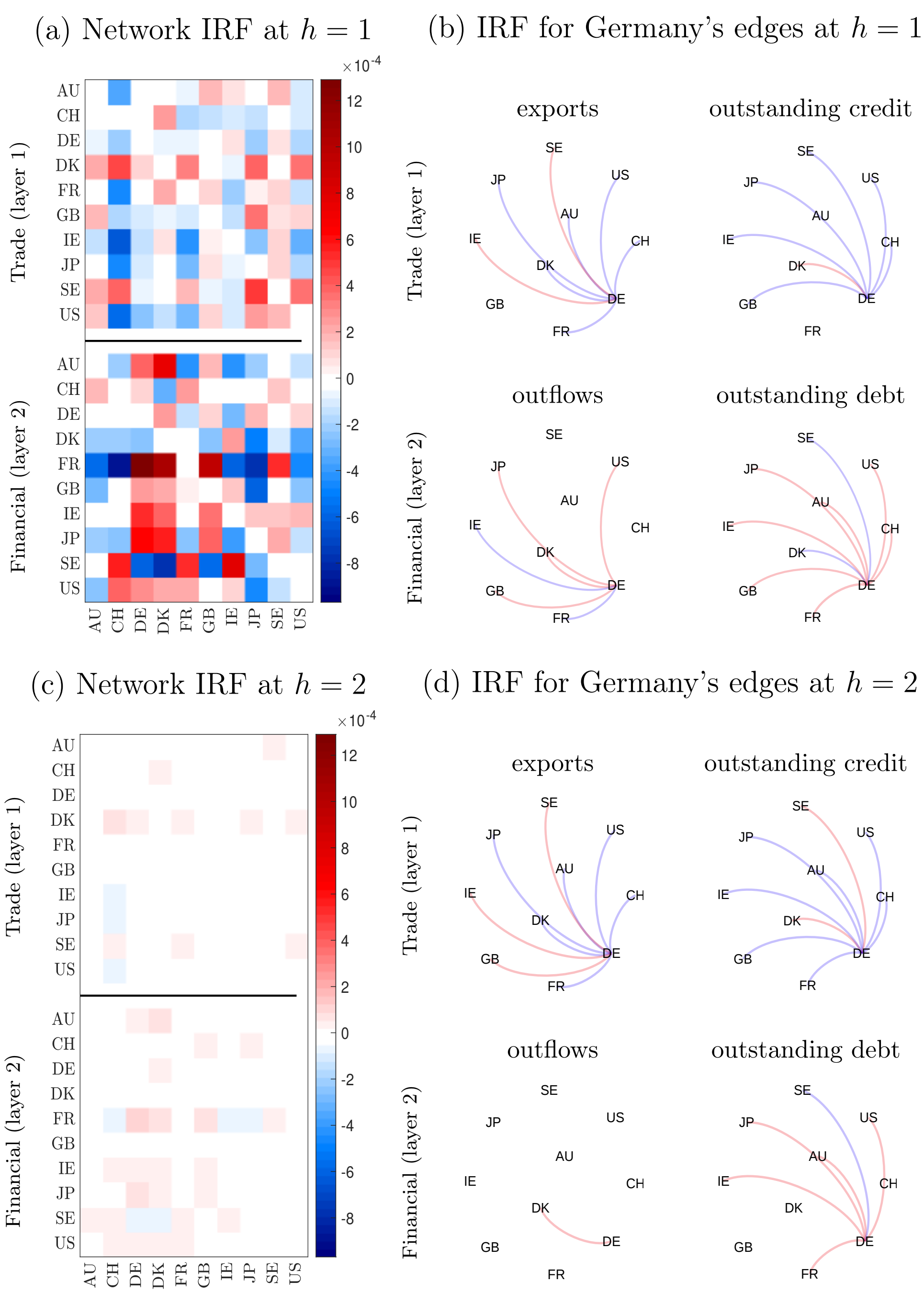

Figure 17: Shock to GB capital inflows by -1%. IRF at horizon $h = 1$ for all (*panel a*) and Germany (*panel b*) financial and trade transactions. IRF at horizon $h = 2$ for all (*panel c*) and Germany (*panel d*) financial and trade transactions. In each plot negative coefficients are in blue and positive in red.

Global effect on the network. We observe heterogeneous effects across countries. Effects on the trade layer at horizon 1 are equally heterogeneous, but smaller in magnitude

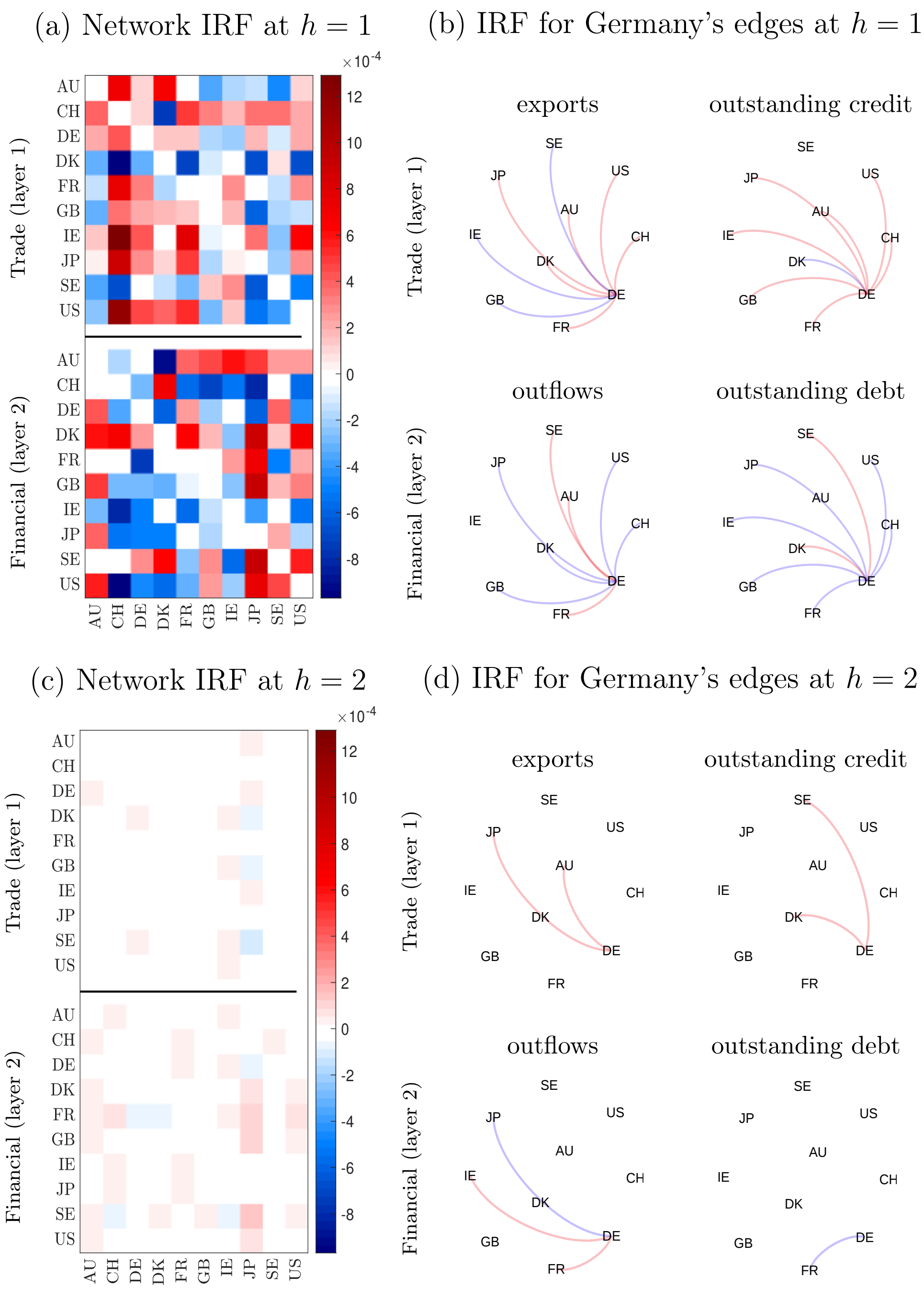

Figure 18: Shock to GB capital inflows by -1% and outflows by +1%. IRF at horizon $h = 1$ for all (*panel a*) and Germany (*panel b*) financial and trade transactions. IRF at horizon $h = 2$ for all (*panel c*) and Germany (*panel d*) financial and trade transactions. In each plot negative coefficients are in blue and positive in red.

compared with the financial layer.

Local effect on Germany. Compared with other countries, the shock has smaller effects on Germany's trade. The negative shock to GB's outstanding debt has a negative impact on Germany's exports and imports to all countries but Ireland and Sweden for exports and Denmark for imports. Germany's outstanding credit increases vis-à-vis Denmark, GB, Japan and US. Germany's outstanding debt increases against all countries but Denmark and Sweden, in particular against France, Japan and Ireland. At horizon 2 responses are not reverted, but nearly all effects turn insignificant, providing evidence of monotone and fast decay of the IRFs.

Local effect on other countries. On the trade layer at horizon 1, we observe a positive response in Denmark's exports and on average a negative response of Switzerland's, Ireland's and Japan's exports. France and Sweden are the most affected countries on the financial layer: The increase in outstanding credit of France towards Germany, Denmark and GB is counterbalanced by a reduction in Sweden's outstanding credit towards the same countries. We observe reverse effects concerning France's and Sweden's outstanding credit towards Switzerland and Ireland. Finally, Ireland's outstanding credit reacts positively towards most other countries.

Compared with responses to the shock to US imports, the persistence of a negative shock to GB's outstanding debt is slightly stronger, see impulse responses at horizon 2 in Fig. 17. The decay is monotonic. However, the speed of decay is heterogeneous across countries. For some countries, there are small effects at horizon 2, while for others the effects are completely wiped already. Overall, we do not find evidence of a relation between the size of a country in terms of exports or outstanding credit and the persistence in the impulse response. At the most, persistence seems determined by the origin of the shock, the effects of a financial shock being more persistent than those of a trade shock.

Finally, in Fig. 18 we plot the block Cholesky IRF, respectively, at horizon $h = 1, 2$, resulting from a 1% negative shock to GB's outstanding debt coupled with a 1% positive shock to GB's outstanding credit. The main findings follow.

Global effect on the network. The results remarkably differ from the previous ones (see Fig. 17). The responses to this simultaneous shock in GB's outstanding debt and credit are larger, in particular in the trade layer. However, already at horizon 2 responses are nearly fully decayed. The results in Fig. 17 and Fig. 18 suggest that an increase in

GB's outstanding credit has an overall positive effect on trade, stimulating export/import activities of most other countries.

Local effect on Germany. One period after the shock, we observe an overall positive effect on German exports, the exception being towards GB, Ireland and Sweden. Imports react mostly positively. Imports from US and Ireland react most, while those from Denmark react negatively. The responses of Germany's outstanding debt vis-à-vis most countries but Denmark and Sweden are negative, especially against France. At horizon 2 Germany's responses have nearly faded away, suggesting a rapid monotone decay of the shock's effect.

Local effect on other countries. In particular, the reactions of Switzerland's imports and outstanding debt are strikingly different from the previous case, compare with Fig. 17. Imports from US and Ireland, and to a lesser extent from France and Austria, are strongly boosted, while those from Denmark and Sweden decrease strongly. Moreover, we note that Japan's outstanding debt increases significantly against most countries. We interpret this as a signal for Japan's attractiveness for foreign capital. Compared with the previous exercise, France's financial responses are now mostly insignificant, or of opposite sign. Finally, the reactions of GB's exports and outstanding credit are heterogeneous, the latter ones being larger in absolute magnitude.

References

Anundsen, A. K., K. Gerdrup, F. Hansen, and K. Kragh-Sørensen (2016). Bubbles and crises: The role of house prices and credit. *Journal of Applied Econometrics 31*(7), 1291–1311.

Baştürk, N., L. Hoogerheide, and H. K. van Dijk (2017). Bayesian analysis of boundary and near-boundary evidence in econometric models with reduced rank. *Bayesian Analysis 12*(3), 879–917.

Behera, R., A. K. Nandi, and J. K. Sahoo (2020). Further results on the Drazin inverse of even order tensors. *Numerical Linear Algebra with Applications 27*(5), e2317.

Brazell, M., N. Li, C. Navasca, and C. Tamon (2013). Solving multilinear systems via tensor inversion. *SIAM Journal on Matrix Analysis and Applications 34*(2), 542–570.

Canova, F. and M. Ciccarelli (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *Journal of Econometrics 120*(2), 327–359.

Canova, F. and M. Ciccarelli (2009). Estimating multicountry VAR models. *International Economic Review 50*(3), 929–959.

Canova, F., M. Ciccarelli, and E. Ortega (2007). Similarities and convergence in G-7 cycles. *Journal of Monetary Economics 54*(3), 850–878.

Carriero, A., G. Kapetanios, and M. Marcellino (2016). Structural analysis with multivariate autoregressive index models. *Journal of Econometrics 192*(2), 332 – 348.

Carroll, J. D. and J.-J. Chang (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrica 35*, 283–319.

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

Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterington (2006). Deviance information criteria for missing data models. *Bayesian analysis 1*(4), 651–673.

Eaton, J. and S. Kortum (2002). Technology, geography, and trade. *Econometrica 70*(5), 1741–1779.

Engle, R. F. and C. W. Granger (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 251–276.

Fagiolo, G., J. Reyes, and S. Schiavo (2009). World-trade web: Topological properties, dynamics, and evolution. *Physical Review E 79*(3), 036115.

Fieler, A. C. (2011). Nonhomotheticity and bilateral trade: Evidence and a quantitative explanation. *Econometrica 79*(4), 1069–1101.

Geweke, J. (1992). Evaluating the accurating of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, pp. 169–193. Oxford University Press.

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017). Bayesian tensor regression. *Journal of Machine Learning Research 18*(79), 1–31.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics, 16, 1- 84*.

Hidalgo, C. A. and R. Hausmann (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences 106*(26), 10570–10575.

Kharrazi, A., E. Rovenskaya, and B. D. Fath (2017). Network structure impacts global commodity trade growth and resilience. *PloS One 12*(2), e0171184.

Kiers, H. A. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chameometrics 14*(3), 105–122.

Kolda, T. G. (2006). Multilinear operators for higher-order decompositions. Technical report, Sandia National Laboratories.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*(3), 455–500.

Koop, G., M. H. Pesaran, and S. M. Potter (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics 74*(1), 119–147.

Kruijer, W., J. Rousseau, and A. Van Der Vaart (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics 4*, 1225–1257.

Lanne, M. and H. Nyberg (2016). Generalized forecast error variance decomposition for linear and nonlinear multivariate models. *Oxford Bulletin of Economics and Statistics 78*(4), 595–603.

LeSage, J. P. (1999). Applied econometrics using matlab. *Manuscript, Dept. of Economics, University of Toronto*, 154–159.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Meyfroidt, P., T. K. Rudel, and E. F. Lambin (2010). Forest transitions, trade, and the global displacement of land use. *Proceedings of the National Academy of Sciences 107*(49), 20917–20922.

Pan, R. (2014). Tensor transpose and its properties. *arXiv preprint arXiv:1411.1503*.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2007). *Numerical recipes: The art of scientic computing*. Cambridge University Press.

Raftery, A. E. and S. M. Lewis (1995). The number of iterations, convergence diagnostics and generic metropolis algorithms. *Practical Markov Chain Monte Carlo 7*(98), 763–773.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods*. Springer.

Schotman, P. and H. K. Van Dijk (1991). A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics 49*(1-2), 195–238.

Schweitzer, F., G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White (2009). Economic networks: The new challenges. *Science 325*(5939), 422–425.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Springer, M. and W. Thompson (1970). The distribution of products of beta, gamma and Gaussian random variables. *SIAM Journal on Applied Mathematics 18*(4), 721–737.

Squartini, T., G. Fagiolo, and D. Garlaschelli (2011). Randomizing world trade. I. A binary network analysis. *Physical Review E 84*(4), 046117.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association 57*(298), 348–368.

Zhu, Z., F. Cerina, A. Chessa, G. Caldarelli, and M. Riccaboni (2014). The rise of China in the international trade network: A community core detection approach. *PloS One 9*(8), e105496.