Bayesian Dynamic Tensor Regression

Monica Billio[†], **Roberto Casarin**[†], Sylvia Kaufmann[‡], Matteo Iacopini[†]

[†] Ca' Foscari University of Venice, Italy

[‡] Study Center Gerzensee, Foundation of the Swiss National Bank, Switzerland

NBP Workshop on Forecasting

26th November 2018

STUDY CENTER
GERZENSEE

Motivation

Availability of data:

(i) increasing size $\Rightarrow$ high dimensionality

(ii) multiple data sources $\Rightarrow$ multiple "layers" (e.g., cross section, time, space, ...)

$\Rightarrow$ gathered or meaningfully rearranged into multidimensional arrays (tensors).

Example 1.

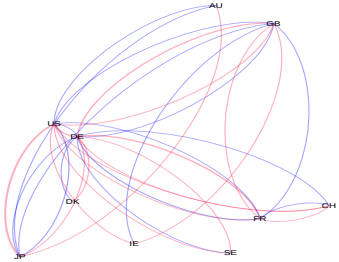
Tensor-valued data:

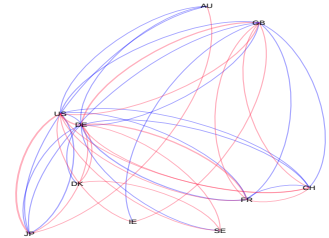
- multi-country panel: m variables, n countries, t times $\rightarrow$ 3-order tensor (e.g., Hoff (2015), Canova and Ciccarelli (2004)).
- temporal networks: relations between n subjects, observed t times $\rightarrow$ 3-order tensor (e.g., financial networks Billio et al. (2012)).
- medical data: sequence of $n \times m$ brain images $\rightarrow$ 3-order tensor (e.g., Zhou et al. (2013), Li and Zhang (2017)).
- **multi-layer networks**: relations between n subjects, d attributes, observed t times $\rightarrow$ 4-order tensor (e.g., social networks Hoff et al. (2002))

Motivation: COMTRADE & BIS Multi-Layer Networks

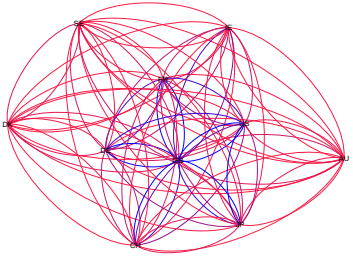
Layer/Time (2004) (2016)

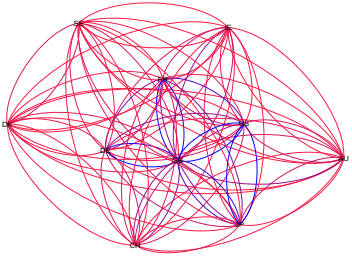
Trade (layer 1)





Financial (layer 2)





Motivation: COMTRADE & BIS Multi-Layer Networks

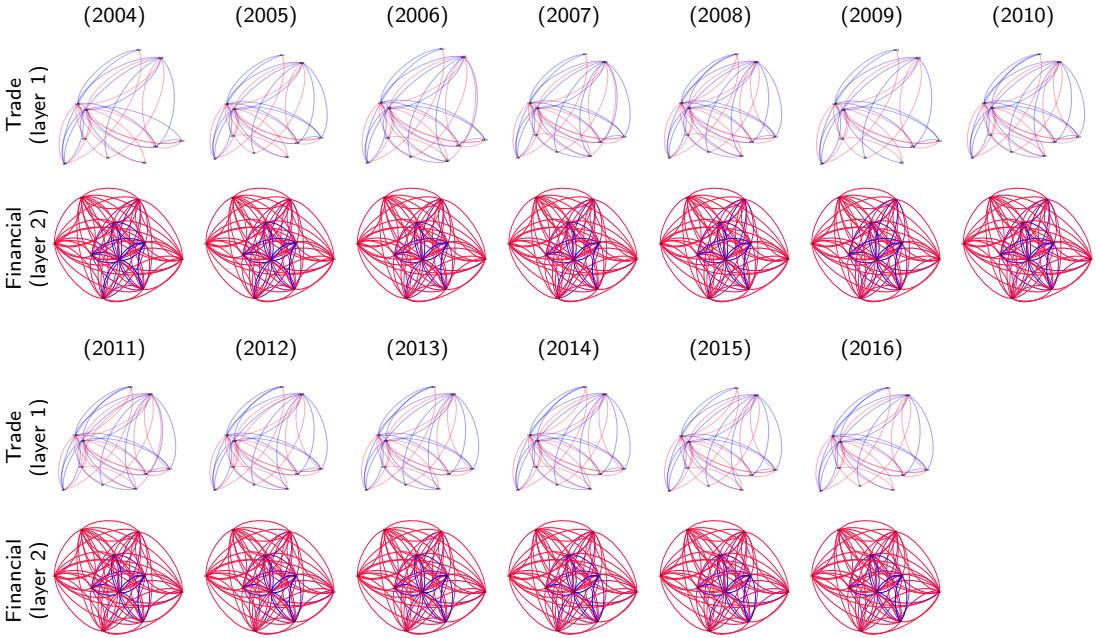


Figure: International trade and financial **temporal networks**. Nodes: countries. Edges: flows.

Questions and Aims

Research questions:

Q: how to model a time series of tensor-valued data?

Q: many variables, few relevant $\Rightarrow$ how to account for sparsity?

Q: possible to exploit information from the structure of the data?

Goals:

G: provide a dynamic model for tensor-valued data

G: explore dynamics of (shock propagation) on tensors

Our contribution:

C: use tensors algebra (spaces, operations and representations)

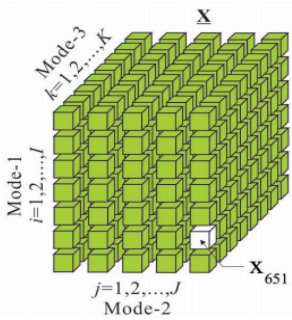
C: use global-local hierarchical prior distributions (information sharing, sparsity)

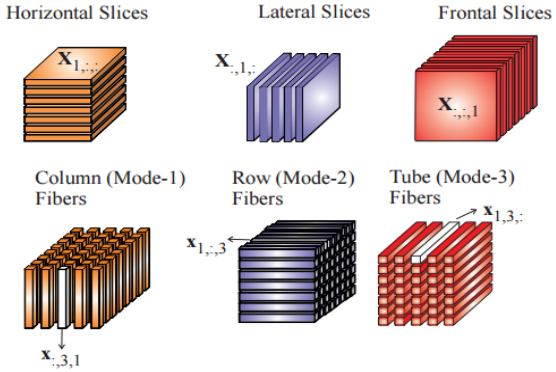
C: extend to tensor models the impulse response analysis

Tensors

Definition 1 (Tensor).

A real valued order-D tensor is an array $\mathcal{X} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$.





✓ Tensor algebra generalizes matrix algebra to multiple dimensions

Tensors Operations

Definition 2 (mode-n product).

Let $\mathcal{X} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$ be a order-D tensor, $A \in \mathbb{R}^{J \times d_n}$ and $\mathbf{v} \in \mathbb{R}^{d_n}$. The mode$-n$ product $\times_n$ is defined as follows:

$$(\mathcal{X} \times_n A)_{i_1,\ldots,i_{n-1},j,i_{n+1},\ldots,i_D} := \sum_{i_n=1}^{d_n} x_{i_1,\ldots,i_n,\ldots,i_D} a_{j,i_n}$$

$$(\mathcal{X} \times_n \mathbf{v})_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_D} := \sum_{i_n=1}^{d_n} x_{i_1,\ldots,i_n,\ldots,i_D} v_{i_n}$$

Idea: compute the inner product of each mode-n fiber with the matrix/vector.

Effect: change n-th dimension of the tensor or reduces its order by one.

- Some operations performed in usual way: scalar product, vectorization, Hadamard, ...
- See also Kolda and Bader (2009), Cichocki et al. (2016).

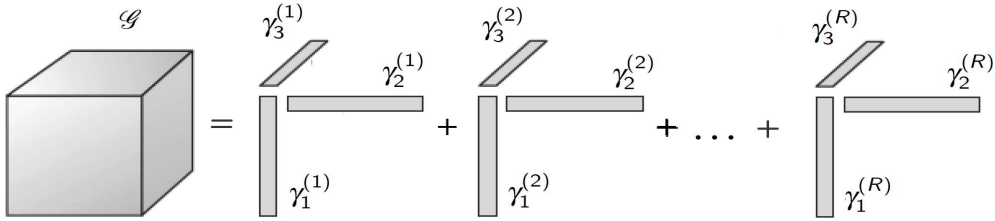
Tensors Representations

Powerful tool: several **tensor representations/decompositions** available (Tucker, PARAFAC, . . .)

Definition 3 (PARAFAC decomposition).

Let $\mathcal{G} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$ and let $R \in \mathbb{R}$ be the rank of $\mathcal{G}$. It holds:

$$\mathcal{G} = \sum_{r=1}^{R} \gamma_1^{(r)} \circ \ldots \circ \gamma_D^{(r)} \qquad \gamma_j^{(r)} \in \mathbb{R}^{d_j}. \tag{1}$$



Remark: R fixed $\Rightarrow$ model selection problem.

A Tensor Model

Tensor Regression

For each entry of the response tensor:

$$y_{\mathbf{i},t} = \beta'_{\mathbf{i}} \operatorname{vec}(\mathcal{X}_t) + \epsilon_{\mathbf{i},t}, \tag{2}$$

where $\mathbf{i} := (i_1, \ldots, i_D)$. Compactly:

$$\begin{aligned} \mathcal{Y}_t &= \mathcal{B} \times_{D+1} \operatorname{vec}(\mathcal{X}_t) + \mathcal{E}_t \\ \mathcal{E}_t &\sim \mathcal{N}_{I_1,\ldots,I_D}(\mathbf{0}, \Sigma_1, \ldots, \Sigma_D) \end{aligned} \tag{3}$$

- $\mathcal{Y}_t, \mathcal{X}_t$: response and regressor **tensors**, with possibly <u>different order</u> and/or size
- $\mathcal{B}$: coefficient **tensor**, size $I_1 \times \ldots \times I_D \times (\prod_j J_j)$
- $\mathcal{E}_t$ noise, with tensor Normal distribution (see Ohlson et al. (2013))
- straightforward inclusion of other regressors: scalars, vectors, matrices, ...

Existing Special Cases

Univariate regression

If $d_i = 1 \, \forall i \in [1, \ldots, D]$ then model (3) reduces to:

$$y_t = \boldsymbol{\beta}' \operatorname{vec}(\mathcal{X}_t) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

Multivariate regression

If $d_i = 1 \; \forall \, i = 2, \ldots, D$ then model (3) reduces to:

$$\mathbf{y}_t = \mathbf{B} \times_2 \operatorname{vec}(\mathcal{X}_t) + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_{l_1}(\mathbf{0}, \Sigma) \tag{5}$$

Sub-cases:

- SUR, when $W_t = 0$, $\mathcal{X}_t = (I_{nm} \otimes X)$ with $X = [X_1, \ldots, X_n]$, $X_i \in \mathbb{R}^{m \times k_i}$, $\mathbf{y}_t = \mathbf{y} \in \mathbb{R}^{nm}$
- VARX(1), when $W_t = 0$, $\mathcal{X}_t = \mathbf{y}_{t-1}$
- Panel VAR(1), when $W_t = 0$, $\mathbf{y}_t = [\mathbf{y}_{1t}, \mathbf{y}_{2t}]$ and $\mathbf{x}_t = g(\mathbf{y}_{t-1})$

New Special Cases

Matrix (dynamic) regression model

Consider a particular case of model (3):

$$\begin{aligned} \mathbf{Y}_t &= \mathcal{B} \times_3 \text{vec}(\mathbf{Y}_t) + \mathbf{E}_t \\ \mathbf{E}_t &\sim \mathcal{N}_{I,I}(\mathbf{0}, \Sigma_1, \Sigma_2) \end{aligned} \tag{6}$$

where $\mathbf{Y}_t \in \mathbb{R}^{I \times I}$ and $\mathcal{B} \in \mathbb{R}^{I \times I \times I^2}$.

Use of matrix variate models/distributions:

- state space time series models Harrison and West (1999)
- Gaussian graphical models Carvalho et al. (2007)
- dynamic linear models Carvalho and West (2007), Wang and West (2009)
- longitudinal data classification and modelling Viroli (2011), Viroli and Anderlucci (2013)
- matrix regression Viroli (2012), Ding and Cook (2016)

Proposed Parametrization

Idea

Each entry of covariate $\mathcal{Y}_t$ depends linearly on every entry of the regressor $\mathcal{X}_t$.

$\Rightarrow$ use tensor $\mathcal{B}$ and mode-d product

Unrestricted model: estimation **infeasible**.

$\Downarrow$

PARAFAC decomposition for $\mathcal{B}$

$$\boxed{\mathcal{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_D^{(r)}}$$

$\Downarrow$

Restricted model: estimation **feasible**

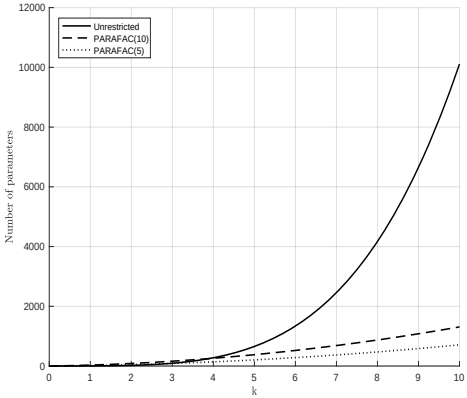


Figure: parameter reduction.

Parametrization Issues

Q: Identification of PARAFAC marginals?

(i) scale invariance

$$\lambda_{1r}\boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \lambda_{Dr}\boldsymbol{\beta}_D^{(r)} = \boldsymbol{\beta}_j^{(r)} \circ \boldsymbol{\beta}_i^{(r)}, \quad \forall\, \lambda_{jr} : \prod_j \lambda_{jr} = 1$$

(ii) permutation invariance

$$\boldsymbol{\beta}_1^{\pi(r)} \circ \ldots \circ \boldsymbol{\beta}_D^{\pi(r)} = \boldsymbol{\beta}_1^{(r)} \circ \ldots \circ \boldsymbol{\beta}_D^{(r)}, \quad \forall \text{ permutation } \pi(\cdot)$$

(iii) (if $D = 2$) invariance up to multiplication by orthonormal vectors

$$\left(\boldsymbol{\beta}_j^{(r)}\mathbf{c}'\right) \circ \left(\boldsymbol{\beta}_i^{(r)}\mathbf{c}'\right) = \boldsymbol{\beta}_j^{(r)} \circ \boldsymbol{\beta}_i^{(r)}, \quad \forall\, \mathbf{c} \in \mathbb{R}^{d_k} : \mathbf{c}'\mathbf{c} = 1$$

Answer $\Rightarrow$ interested in $\mathcal{B}$, always identified <u>not</u> in its marginals $\boldsymbol{\beta}_j^{(r)}$

Prior Specification

Idea

Only a fraction of the regressors have significant effect $\Rightarrow$ sparse coefficient

Bayesian approach + PARAFAC

$\Downarrow$

choice of shrinkage priors on tensor marginals

$\Downarrow$

induced sparsity on the tensor of coefficients $\mathcal{B}$.

Remark 1.

PARAFAC tensor decomposition allows to change the choice of prior distribution:
(rigid) prior on tensor $\mathcal{B} \Rightarrow$ (flexible) prior on vectors $\beta_d^{(r)}$

Prior Specification

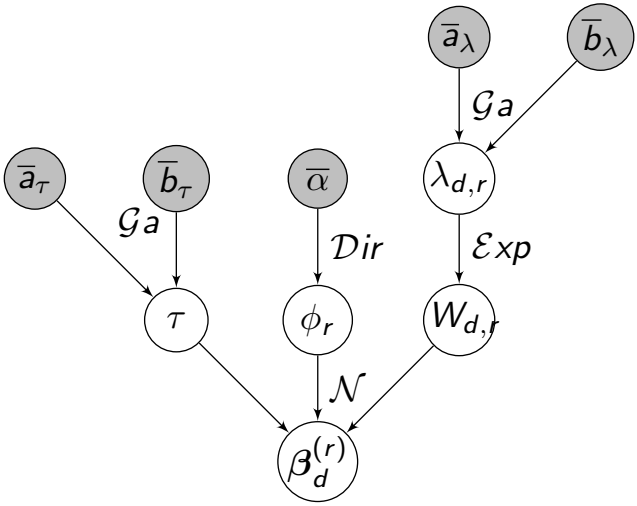


Figure: DAG of hierarchical global-local shrinkage prior.

Prior Specification

Hierarchical **global-local** shrinkage prior for tensor margins:

$$\pi(\boldsymbol{\beta}_d^{(r)}|\tau, \phi_r, W_{d,r}) \sim \mathcal{N}_{I_d}(\mathbf{0}, \underbrace{\tau}_{\text{global}} \underbrace{\phi_r}_{\text{comp}} \underbrace{W_{d,r}}_{\text{local}}) \qquad \forall\, d, r$$

$$\pi(\tau) \sim \mathcal{G}a(\overline{a}_\tau, \overline{b}_\tau), \quad \pi(\phi) \sim \mathcal{D}ir(\overline{\boldsymbol{\alpha}}), \quad \text{global and component parts}$$

$$\pi(\lambda_{d,r}) \sim \mathcal{G}a(\overline{a}_\lambda, \overline{b}_\lambda), \quad \pi(w_{d,r,k}|\lambda_{d,r}) \sim \mathcal{E}xp(\lambda_{d,r}^2/2), \quad \text{local part}$$

$$\pi(\gamma) \sim \mathcal{G}a(\overline{a}_\gamma, \overline{b}_\gamma), \quad \pi(\Sigma_d|\gamma) \sim \mathcal{IW}_{I_d}(\overline{\nu}_d, \gamma\overline{\Psi}_d), \quad \text{noise covariances}$$

Prior Specification Issues

Prior on $\beta_d^{(r)}$

- joint prior on $\mathcal{B}_{ijk}$ is not Normal $\Rightarrow$ heavier tails; distribution
- computationally faster than alternatives (SSVS, spike-and-slab,...)
- identification issue for marginals $\beta_d^{(r)}$: scaling, permutation, multiplication by orthogonal matrix $\Rightarrow$ tensor $\mathcal{B}$ identified

Prior on Σ_d

- conditional prior dependence (via γ) helps avoiding scale un-identification (Hoff (2011))

Posterior Computation

Gibbs sampler in three main steps:
Step 1. sample global and component variance parameters from

$$p(\phi, \tau | \mathcal{B}, \mathbf{W}) = p(\phi | \mathcal{B}, \mathbf{W}) p(\tau | \phi, \mathcal{B}, \mathbf{W})$$

- collapsed Gibbs: $p(\psi_r | \mathcal{B}, \mathbf{W}) \sim \text{GiG}(\alpha - d_0/2, 2b_\tau, 2C_r)$
  then $\phi_r = \psi_r / \sum_l \psi_l \quad \forall\, r$
- $p(\tau | \phi, \mathcal{B}, \mathbf{W}) \sim \text{GiG}(a_\tau - Rd_0/2, 2b_\tau, 2\sum_r D_r)$

Step 2. sample tensor marginals and local variance parameters from

$$p\left(\{\boldsymbol{\beta}_d^{(r)}, \mathbf{w}_{d,r}, \lambda_{d,r}\}_{d,r} | \phi, \tau, \mathbf{Y}, \Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4\right)$$

- $p(\lambda_{d,r} | \phi_r, \tau, \boldsymbol{\beta}_d^{(r)}) \sim \mathcal{G}a\left(a_\lambda + I_d, b_\lambda + \left\|\boldsymbol{\beta}_d^{(r)}\right\|_1 / \sqrt{\tau \phi_r}\right)$
- $p(w_{d,r,k} | \lambda_{d,r}, \phi_r, \tau, \boldsymbol{\beta}_d^{(r)}) \sim \text{GiG}\left(\frac{1}{2}, \lambda_{d,r}^2, \beta_{d,k}^{(r)^2} / (\tau \phi_r)\right) \quad \forall\, k \in [1, I_d]$
- $p(\boldsymbol{\beta}_j^{(r)} | \boldsymbol{\beta}_{-j}^{(r)}, \mathcal{B}_{-r}, \phi, \tau, \mathbf{Y}, \Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4) \sim \mathcal{N}_{I_j}(\boldsymbol{\mu}_{\beta_j}, \Sigma_{\beta_j})$

Posterior Computation

Step 3. sample noise covariance matrices from

$$p(\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4, \gamma | \mathcal{B}, \mathbf{Y}) = p(\gamma | \Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4) \prod_{d=1}^{4} p(\Sigma_d | \gamma, \Sigma_{-d}, \mathcal{B}, \mathbf{Y})$$

- $p(\gamma|-) \sim \mathcal{G}a(\overline{a}_\gamma + (\sum_{d=1}^4 \overline{\nu}_d + TI_d)/2,\ \overline{b}_\gamma + \mathrm{tr}(\sum_{d=1}^4 \overline{\Psi}_d \Sigma_d^{-1})/2)$
- $p(\Sigma_d|\gamma, \Sigma_{-d}, \mathcal{B}, \mathbf{Y}) \sim \mathcal{IW}_{I_d}(\overline{\nu}_d + TI_d,\ \gamma\overline{\Psi}_d + S_d)$

Remark 2 (Init).

- Hamiltonian Monte Carlo steps (Neal (2011)) for improving mixing
- simple to add vector, matrix, tensor covariates

Simulation

Model: matrix regression in eq. (6)

- $\mathbf{Y}_t, \mathbf{X}_t \in \mathbb{R}^{k \times k}$ square matrices
- $\mathcal{B} \in \mathbb{R}^{k \times k \times k^2}$ 3-order coefficient **tensor**
- main parameters:

$$T = 60 \quad x_{ij,t} \sim \text{AR}(1) \text{ not stationary}$$

$$R = 5 \quad \boldsymbol{\beta}_d^{(r)} \overset{iid}{\sim} \mathcal{N}_{I_d}(\boldsymbol{\mu}_d, \boldsymbol{\Omega}_{d,r})$$

Settings:

size $\mathbf{Y}_t, \mathbf{X}_t$	size $\mathcal{B}$	# parameters
10×10	$10 \times 10 \times 100$	$15 \cdot 120$
20×20	$20 \times 20 \times 400$	$15 \cdot 440$
30×30	$30 \times 30 \times 900$	$15 \cdot 960$
40×40	$40 \times 40 \times 1600$	$15 \cdot 1680$
50×50	$50 \times 50 \times 2500$	$15 \cdot 2600$

Simulation results

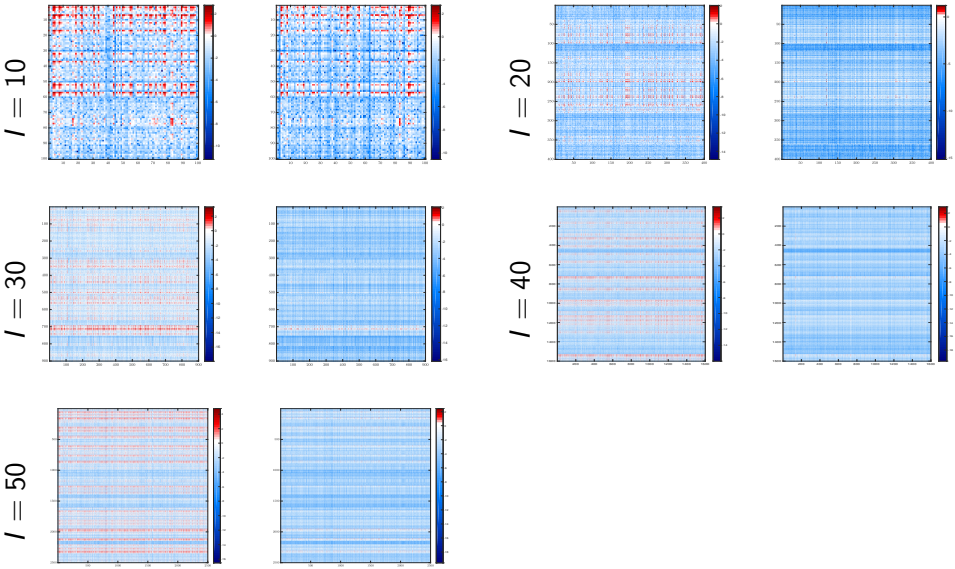


Figure: Log-absolute value of coefficient tensors: true $\mathcal{B}$ (*left*) and estimated $\hat{\mathcal{B}}$ (*right*).

Application I - COMTRADE data

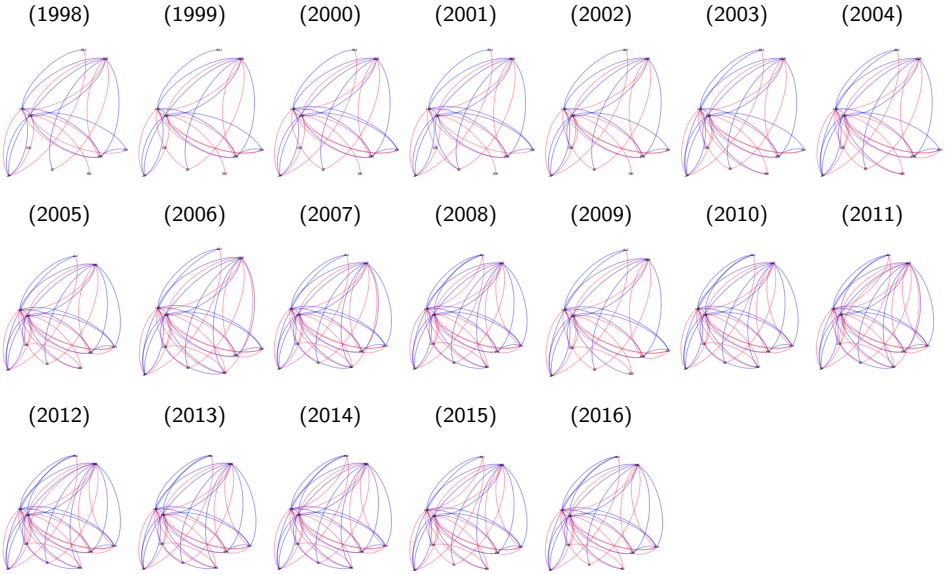


Figure: Trade network from 1998 (*top left*) to 2016 (*bottom right*). Nodes are countries, red and blue edges stand for exports and imports between two countries. Edge thickness represents flow magnitude.

Application I - Coefficient tensor

Matrix autoregressive model

$$\mathbf{Y}_t = \mathcal{B} \times_3 \text{vec}(\mathbf{Y}_{t-1}) + \mathbf{E}_t\,, \qquad \mathbf{E}_t \sim \mathcal{N}_{10,10}(\mathbf{0}, \Sigma_1, \Sigma_2) \tag{7}$$

- mode-3 <u>matricized</u> tensor:

$$\text{mat}_3(\mathcal{B})' = \mathbf{B}'_{(3)} = \left[\text{vec}(\mathcal{B}_{::1}), \text{vec}(\mathcal{B}_{::2}), \ldots, \text{vec}(\mathcal{B}_{::100})\right] \in (10^2 \times 100)$$

- entry (i, j) of $\mathbf{B}'_{(3)}$: impact of edge j $[t-1]$ $\rightarrow$ i $[t]$

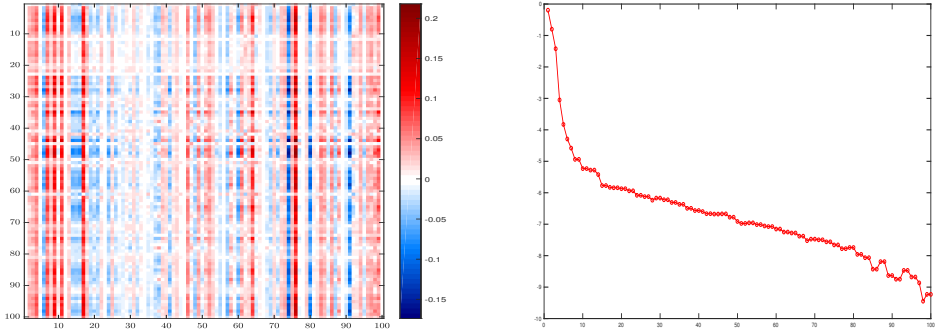


Figure: *Left:* $\hat{\mathbf{B}}'_{(3)}$. *Right:* log-modulus of the eigenvalues of $\hat{\mathbf{B}}_{(3)}$, decreasing order.

Application I - Noise covariances

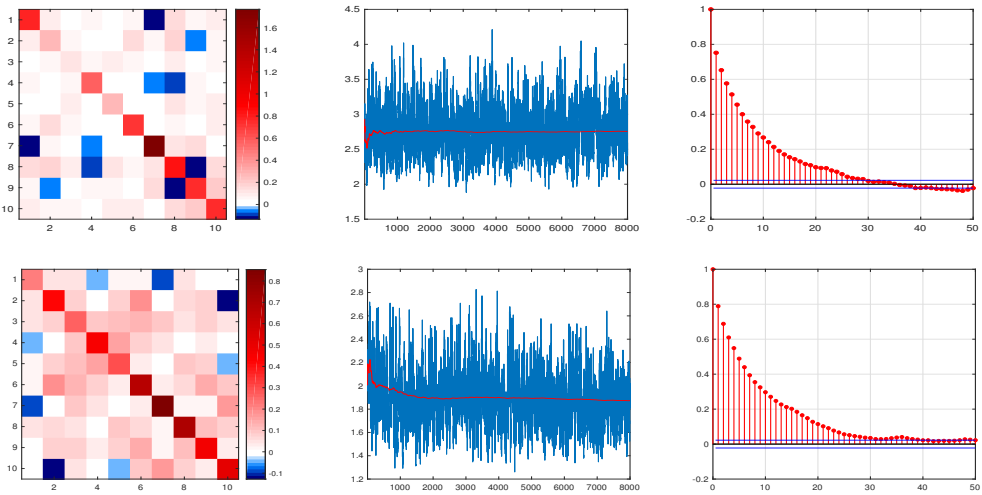


Figure: Estimated covariance matrix of the noise term (*first*), MCMC output (*second*) and autocorrelation functions (*third*) of the Frobenious norm of the covariance matrix of the noise term. *First row*: Σ_1, *second row*: Σ_2.

Impulse Response in Dynamic Tensor Models

Definition 4 (Impulse response function (IRF)).

Let $\widetilde{\mathcal{E}}$ be a tensor of shocks and $\text{vecr}(\cdot)$ the reverse vectorization operator.
The **tensor**-valued **impulse response** function (IRF) at horizon $h \geq 1$ is:

$$\widetilde{\mathcal{Y}}_h = \text{vecr}([\mathbf{B}'_{(D+1)}]^h \cdot \text{vec}(\widetilde{\mathcal{E}}))\,. \tag{8}$$

The IRF for network data allows for:

- temporal + cross-sectional shock propagation analysis
- shock on single/multiple flows (edges) or countries (nodes);
- individual flow level analysis of shock propagation;
- analysis of role of network topology in shock propagation;

Interpreting IRF for tensor

Orthogonalised IRF requires **orthogonal** shocks:

$$IRF_h = \mathbb{E}[\mathcal{Y}_{t+h}|\widetilde{E}_{ij,t} = \delta_{ij}, \widetilde{\mathbf{E}}_{-ij,t} = 0, \mathcal{F}_{t-1}] - \mathbb{E}[\mathcal{Y}_{t+h}|\widetilde{\mathbf{E}}_t = 0, \mathcal{F}_{t-1}] \tag{9}$$

- covariance restrictions for avoiding/mitigating *compositional effect* (due to contemporaneous correlations)
- Cholesky $\Rightarrow$ not invariant to *ordering* of variables; not unique

Generalised IRF (Koop et al. (1996), Pesaran and Shin (1998)):

$$GIRF_h = \mathbb{E}[\mathcal{Y}_{t+h}|\widetilde{E}_{ij,t} = \delta_{ij}, \mathcal{F}_{t-1}] - \mathbb{E}[\mathcal{Y}_{t+h}|\mathcal{F}_{t-1}] \tag{10}$$

- unique and invariant to ordering of variables
- no covariance restrictions: when one variable is shocked, other variables also vary, then average by integrating out all other shocks
- not distinguish causes of a change in $\mathbb{E}[\mathcal{Y}_{t+h}|\mathcal{F}_{t-1}]$

Impulse Response Analysis

Definition 2 (Block-orthogonalized IRF for tensor models).

Denote $\mathbf{\Sigma}$ the covariance matrix of the vectorised Tensor autoregressive model. We propose to use the impulse response function resulting from the transformation

$$\begin{aligned} \mathrm{vec}(\mathcal{Y}_t) &= \sum_{i=0}^{\infty} \mathbf{\Phi}_i \boldsymbol{\epsilon}_{t-i} = \sum_{i=0}^{\infty} (\mathbf{\Phi}_i \mathbf{L})(\mathbf{L}^{-1}\boldsymbol{\epsilon}_{t-i}) \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \\ &= \sum_{i=0}^{\infty} (\mathbf{\Phi}_h \mathbf{L}) \boldsymbol{\eta}_{t-i} \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \end{aligned} \tag{11}$$

where

$$\mathbf{D} = \mathbf{L}^{-1} \cdot \mathbf{\Sigma} \cdot (\mathbf{L}')^{-1} = \left[\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{S} \end{array}\right], \qquad \mathbf{\Phi}_0 = I, \quad \mathbf{\Phi}_i = \mathbf{B}'_{(4)} \mathbf{\Phi}_{i-1}. \tag{12}$$

Application I - Impulse response for nodes

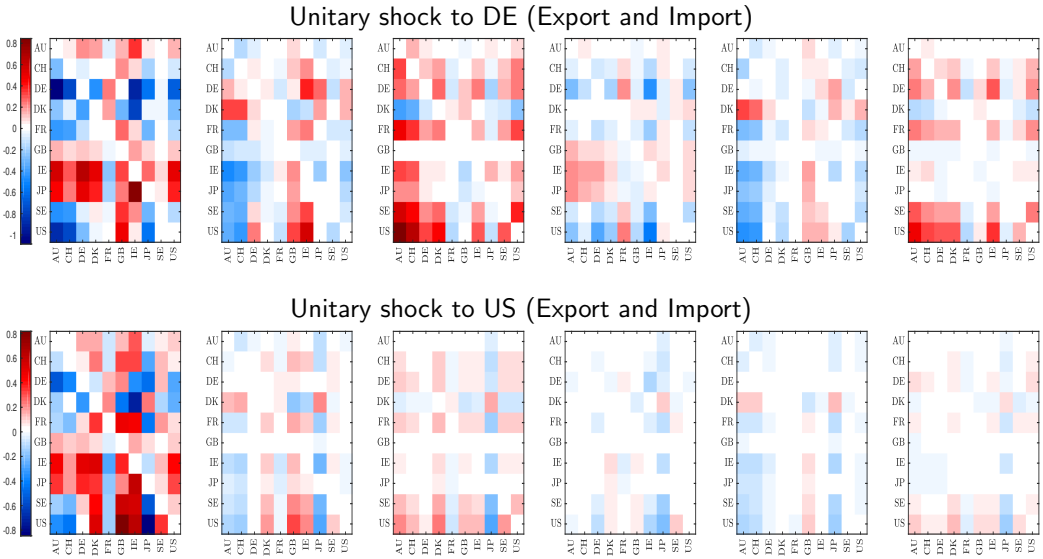


Figure: Positive effects in red, negative effects in blue.

Application II: COMTRADE & BIS Multi-Layer Networks

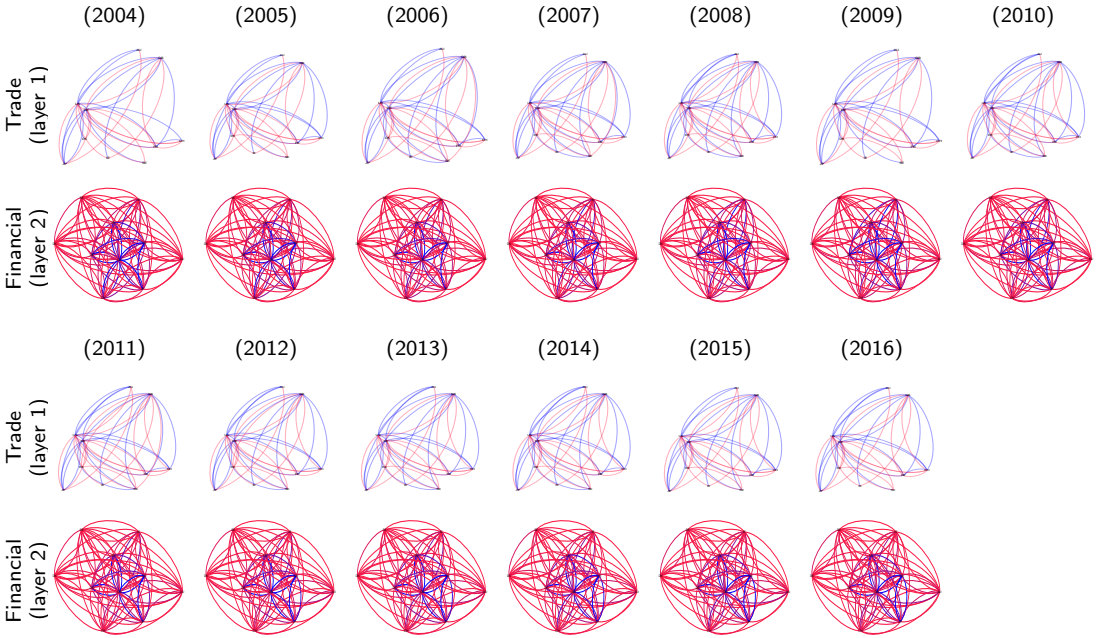


Figure: International trade and financial networks. Nodes: countries. Edges: flows.

Application II - COMTRADE + BIS

Tensor autoregressive model

$$\mathcal{Y}_t = \mathcal{B} \times_4 \text{vec}(\mathcal{Y}_{t-1}) + \mathcal{E}_t$$
$$\mathcal{E}_t \sim \mathcal{N}_{10,10,2}(\mathbf{0}, \Sigma_1, \Sigma_2, \Sigma_3) \tag{13}$$

- mode-3 : 4 <u>matricized</u> tensor:

$$\text{mat}_{3:4}(\mathcal{B})' = \mathbf{B}'_{(3:4)} = \left[\text{vec}(\mathcal{B}_{::1,1}), \ldots, \text{vec}(\mathcal{B}_{::1,200}), \text{vec}(\mathcal{B}_{::2,1}), \ldots, \text{vec}(\mathcal{B}_{::2,200})\right]$$

- entry (i, j) of $\mathbf{B}'_{(3:4)}$: impact of edge j $[t-1]$ $\rightarrow$ i $[t]$

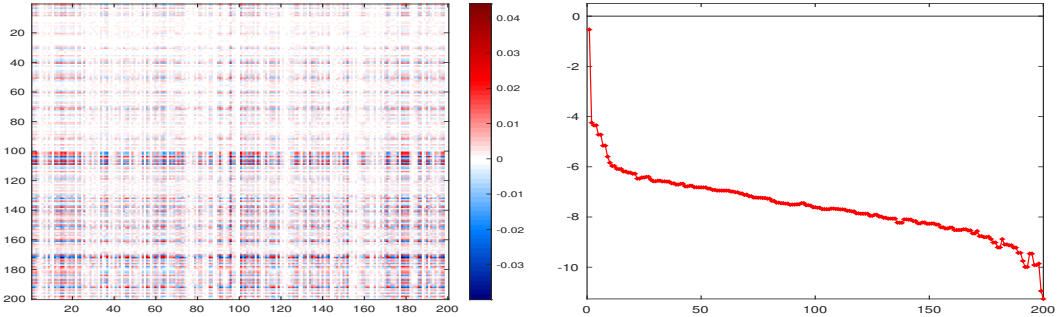
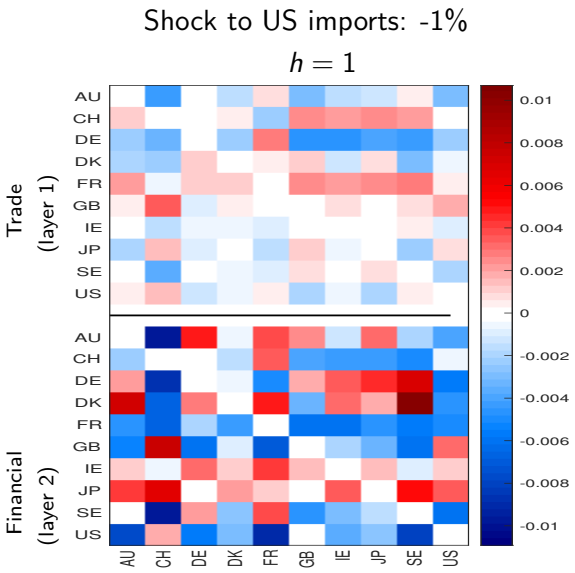


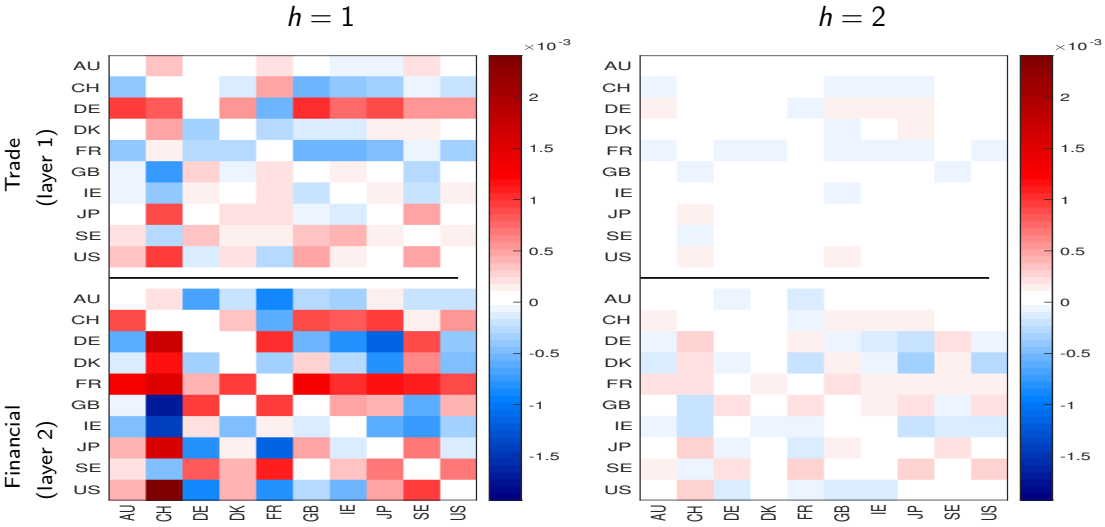
Figure: *Left:* $\hat{\mathbf{B}}'_{(3:4)}$. *Right:* log-modulus of the eigenvalues of $\hat{\mathbf{B}}_{(3:4)}$, decreasing order.

Impulse Response: US import



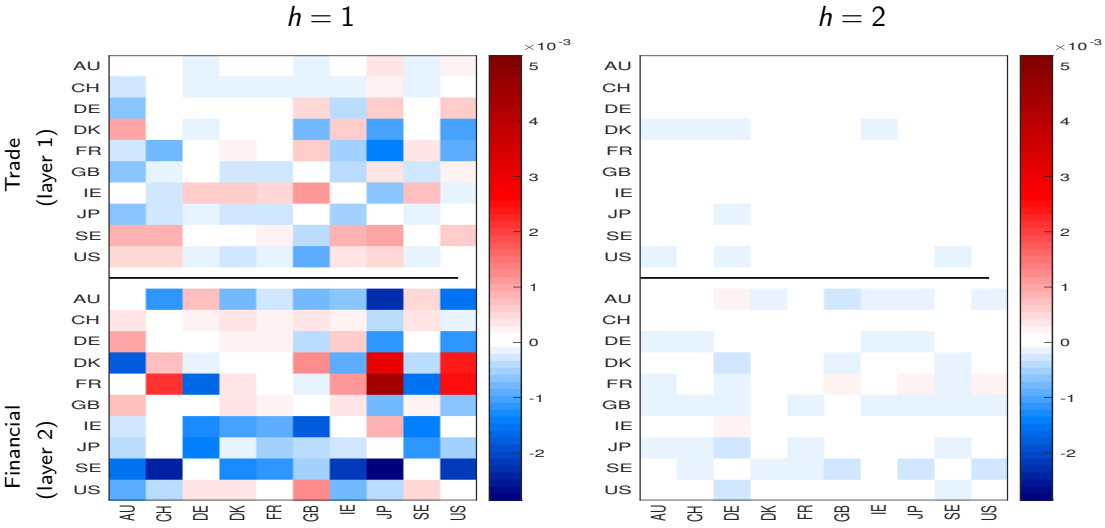
Impulse Response: UK financial flows

Shock to GB capital **inflows: -1%**



Impulse Response: UK financial flows

Shock to GB capital **inflows: -1% and to outflows +1%**



Conclusions

Proposal: linear, dynamic tensor regression model

- generalises linear regression models to multi-dimensional regression
- PARAFAC tensor decomposition for parsimony
- hierarchical global-local shrinkage prior for sparse coefficients
- good performance against synthetic data up to 50×50

- application to COMTRADE network (matrix AR(1) model):
  - ✓ impact of trade links is heterogeneous and sparse
  - ✓ heterogeneous magnitude and persistence of shock propagation
  - ✓ role of network topology in shock propagation
- application to COMTRADE+BIS networks (tensor AR(1) model):
  - ✓ impact of trade and financial links are heterogeneous and sparse
  - ✓ financial shock propagation has higher magnitude
  - ✓ between layer propagation
  - ✓ meaningful country-specific IRF results

Comments

Thank you!

Suggestions/comments are welcome!

Working Paper

Billio, M., Casarin, R., Kaufmann, S. & Iacopini, M. (2018), "*Bayesian Dynamic Tensor Regression*", arXiv preprint 1709.09606.

Billio, M., R. Casarin, S. Kaufmann, and M. Iacopini (2018): "Bayesian Dynamic Tensor Regression," *arXiv preprint arXiv:1709.09606*.

Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon (2012): "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *Journal of Financial Economics*, 104, 535–559.

Bussière, M., A. Chudik, and G. Sestieri (2012): "Modelling Global Trade Flows: Results from a GVAR Model," Tech. Rep. 119, Federal Reserve Bank of Dallas.

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

Carvalho, C. M., H. Massam, and M. West (2007): "Simulation of hyper-inverse Wishart distributions in graphical models," *Biometrika*, 94, 647–659.

Carvalho, C. M. and M. West (2007): "Dynamic matrix-variate graphical models," *Bayesian Analysis*, 2, 69–97.

Cichocki, A., N. Lee, I. Oseledets, A. Phan, Q. Zhao, and D. Mandic (2016): "Low-rank tensor networks for dimensionality reduction and large-scale optimization problems: perspectives and challenges PART 1," *arXiv preprint arXiv:1609.00893*.

Ding, S. and R. D. Cook (2016): "Matrix-variate regressions and envelope models," *arXiv preprint arXiv:1605.01485*.

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

Harrison, J. and M. West (1999): *Bayesian forecasting & dynamic models*, Springer.

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

——— (2015): "Multilinear tensor regression for longitudinal relational data," *The Annals of Applied Statistics*, 9, 1169–1193.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002): "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

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

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

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

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

Neal, R. M. (2011): "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, ed. by S. Brooks, A. Gelman, J. L. Galin, and X.-L. Meng, Chapman & Hall /CRC, chap. 5.

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.

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

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

Viroli, C. (2011): "Finite mixtures of matrix normal distributions for classifying three-way data," *Statistics and Computing*, 21, 511–522.

——— (2012): "On matrix-variate regression analysis," *Journal of Multivariate Analysis*, 111, 296–309.

Viroli, C. and L. Anderlucci (2013): "Modelling longitudinal data through matrix-variate normal mixtures," in *Advances in Latent Variables - Methods, Models and Applications*.

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

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

Prior for entry of tensor $\mathcal{B}$ back

The joint distribution of PARAFAC marginal entries is $\prod_{h=1}^{3} \pi(\beta_{h,i}^{(r)}|\tau, \phi_r, w_{h,r})$. Intuition: to obtain the distribution of tensor entry b_{ijk} apply:

Theorem 3 (4 in Springer and Thompson (1970)).

The probability density function of the product $z = \prod_{j=1}^{J} x_j$ of J independent Normal random variables $x_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, \ldots, J$, is a Meijer G-function multiplied by a normalising constant H:

$$p(z|0, \{\sigma_j^2\}_{j=1}^{J}) = H \cdot G_{J,0}^{J,0}\left(z^2 \cdot \prod_{j=1}^{J} \frac{1}{2\sigma_j} \middle| \mathbf{0}\right), \tag{14}$$

where

$$H = (2\pi)^{-J/2} \cdot \prod_{j=1}^{J} \sigma_j^{-1} \tag{15}$$

and $G_{p,q}^{m,n}(\cdot|\cdot)$ is a Meijer G-function ($c \in \mathbb{R}, s \in \mathbb{C}$, integral along vertical line in the complex plane):

$$G_{p,q}^{m,n}\left(z \middle| \begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix}\right) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} z^{-s} \frac{\prod_{j=1}^{m} \Gamma(s + b_j) \cdot \prod_{j=1}^{n} \Gamma(1 - a_j - s)}{\prod_{j=n+1}^{p} \Gamma(s + a_j) \cdot \prod_{j=m+1}^{q} \Gamma(1 - b_j - s)} \, \mathrm{d}s \,. \tag{16}$$

Prior for entry of tensor $\mathcal{B}$

Marginal prior distribution of generic tensor entry:

$$\pi(b_{ijk}) = \int \pi(b_{ijk}|\tau, \boldsymbol{\phi}, \mathbf{W})\pi(\tau)\pi(\boldsymbol{\phi})\pi(\mathbf{W})\,\mathrm{d}\tau\,\mathrm{d}\boldsymbol{\phi}\,\mathrm{d}\mathbf{W}\,, \tag{17}$$

where, denoting $\beta_r = \beta_{1,i}^{(r)} \cdot \beta_{2,j}^{(r)} \cdot \beta_{3,k}^{(r)}$:

$$\pi(b_{ijk}|\tau, \boldsymbol{\phi}, \mathbf{W}) = p\left(\sum_{r=1}^{R} H_r \cdot G_{3,0}^{3,0}\left(\beta_r^2 \cdot \prod_{h=1}^{3} \frac{1}{2\tau\phi_r} W_{h,r}^{-1}\middle|\mathbf{0}\right)\right)$$

with:

$$G_{3,0}^{3,0}\left(\beta_r^2 \cdot \prod_{h=1}^{3} \frac{1}{2\tau\phi_r} W_{h,r}^{-1}\middle|\mathbf{0}\right) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \left(\frac{\beta_r^2}{(2\tau\phi_r)^3}\prod_{h=1}^{3} W_{h,r}^{-1}\right)^{-s} \mathrm{d}s \tag{18}$$

$$H_r = (2\pi)^{-2} \cdot (\tau\phi_r)^{-3} \prod_{j=1}^{3} W_{h,r}^{-1} \tag{19}$$

Prior for entry of tensor $\mathcal{B}$

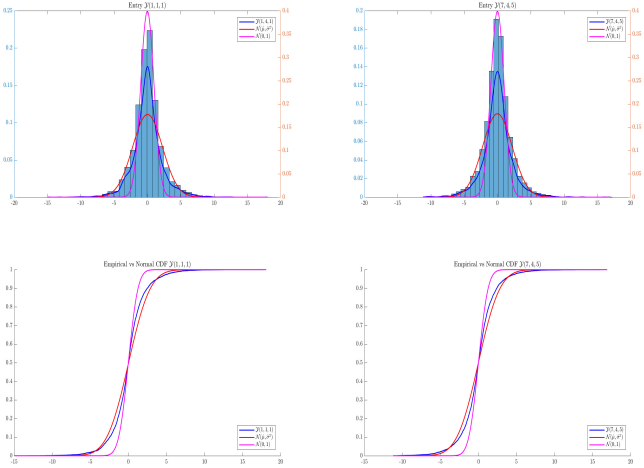


Figure: Simulated distribution of two entries of tensor (std Normal marginals) vs Normal with same mean and variance vs std Normal, for $R = 5$.

Prior for entry of tensor $\mathcal{B}$ back

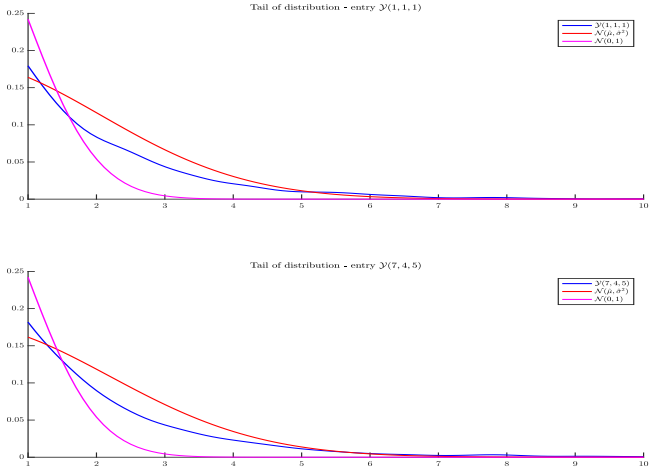


Figure: Simulated distribution (tail) of two entries of tensor (std Normal marginals) vs Normal with same mean and variance vs std Normal, for $R = 5$.

Prior for entry of tensor $\mathcal{B}$

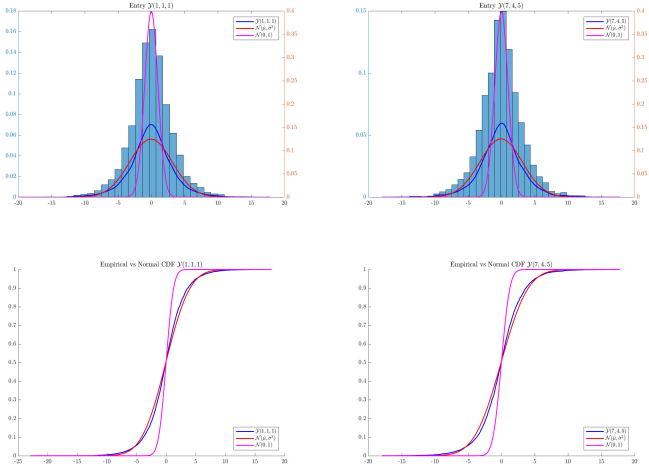


Figure: Simulated distribution of two entries of tensor (std Normal marginals) vs Normal with same mean and variance vs std Normal, for $R = 10$.

Prior for entry of tensor $\mathcal{B}$ back

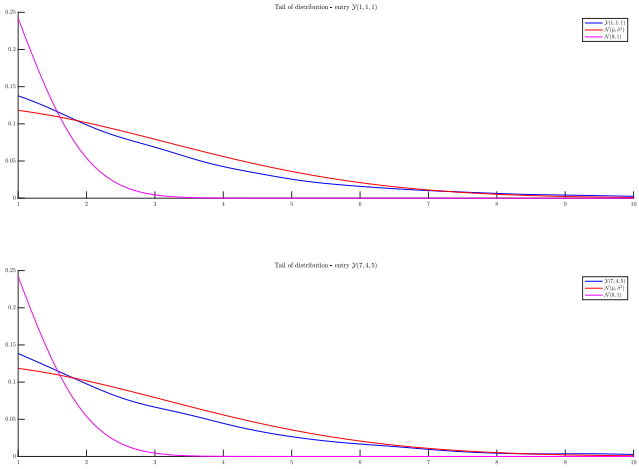


Figure: Simulated distribution (tail) of two entries of tensor (std Normal marginals) vs Normal with same mean and variance vs std Normal, for $R = 10$.

Initialisation of Gibbs sampler

Gibbs sampler sensitive to initial value of some key parameters:

- tensor PARAFAC marginals $\{\beta_j^{(r)}\}_{j,r}$ initialised via Simulated Annealing;
  Intuition: find the set of marginals generating a sufficiently sparse tensor, while allowing deviations from zero.
- other parameters initialised from prior distribution;
  Intuition: sampler not very sensitive to their starting value.

Initialisation marginals $\beta_j^{(r)}$ - back

Intuition: find the set of marginals generating a tensor with many entries close to zero, others far zero. Use Simulated Annealing for minimising the objective function:

$$f(\tilde{\mathcal{B}}^{(n)}) = \psi_0 \left\| \tilde{\mathcal{B}}^{(n)} \right\|_2 + \psi_3 \sum_{r=1}^{R} \left\| \tilde{\boldsymbol{\beta}}_3^{(r),(n)} \right\|_2 . \tag{20}$$

Penalties:

- $\psi_0 > 0 \rightarrow$ tensor quadratic norm;
- $\psi_3 > 0 \rightarrow$ quadratic norm 3-order marginals.

Cooling schedule, with fixed $q > 0$:

$$C(n) = \frac{q}{1 + \log(n)} \qquad n = 1, \ldots, N_{SA} . \tag{21}$$