Università
Ca' Foscari
Venezia

Three Essays on Tensor Regression Models

Qing Wang

Università Ca' Foscari Venezia
Department of Economics

PhD defense

5th March 2026

Motivation

Tensor regression

Linear regression:

$$y_t = \boldsymbol{\beta}^\top \boldsymbol{x}_t + \sigma\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1)$$

where $y_t \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$, $\boldsymbol{x}_t \in \mathbb{R}^d$.

Motivation

Tensor regression

Linear regression:

$$y_t = \boldsymbol{\beta}^\top \boldsymbol{x}_t + \sigma \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1)$$

where $y_t \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$, $\boldsymbol{x}_t \in \mathbb{R}^d$.

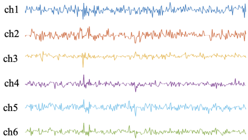
Tensor regression (scalar-on-tensor):

$$y_t = \langle \mathcal{B}, \mathcal{X}_t \rangle + \sigma \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1)$$

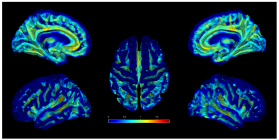
where $\langle , \rangle$ denotes the inner product, $\mathcal{B}, \mathcal{X}_t \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_M}$.

Motivation

Tensor Regression

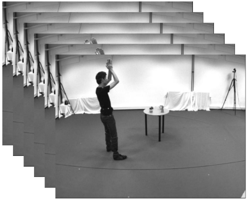


(a) ECoG signals



(b) fMRI images

(c) Facial images



(d) Video sequences

Figure: Liu et al. (2021)

Motivation

Tensor decomposition

Several **tensor representations/decompositions** available (Tucker, PARAFAC, Tensor-Train, etc.)

PARAFAC(R) decomposition

Let $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_M}$ and let $R \in \mathbb{N}$ be the rank of $\mathcal{X}$. It holds:

$$\mathcal{X} = \sum_{r=1}^{R} \gamma_1^{(r)} \circ \dots \circ \gamma_M^{(r)}, \qquad \gamma_m^{(r)} \in \mathbb{R}^{d_m}. \tag{1}$$

Motivation

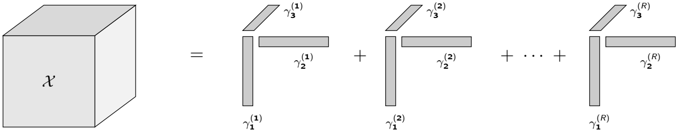
Tensor decomposition

Several **tensor representations/decompositions** available (Tucker, PARAFAC, Tensor-Train, etc.)

PARAFAC(R) decomposition

Let $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_M}$ and let $R \in \mathbb{N}$ be the rank of $\mathcal{X}$. It holds:

$$\mathcal{X} = \sum_{r=1}^{R} \gamma_1^{(r)} \circ \ldots \circ \gamma_M^{(r)}, \qquad \gamma_m^{(r)} \in \mathbb{R}^{d_m}. \tag{1}$$



Motivation

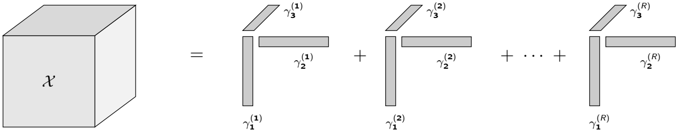
Tensor decomposition

Several **tensor representations/decompositions** available (Tucker, PARAFAC, Tensor-Train, etc.)

PARAFAC(R) decomposition

Let $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_M}$ and let $R \in \mathbb{N}$ be the rank of $\mathcal{X}$. It holds:

$$\mathcal{X} = \sum_{r=1}^{R} \gamma_1^{(r)} \circ \ldots \circ \gamma_M^{(r)}, \qquad \gamma_m^{(r)} \in \mathbb{R}^{d_m}. \tag{1}$$



Change of parameters: $\prod_{m=1}^{M} d_m \to R(\sum_{m=1}^{M} d_m)$

Università Ca' Foscari Venezia

Thesis at a glance

Goal: How can we build flexible, scalable and interpretable Bayesian tensor regression models for high-dimensional multiway data?

Chapter 1: Regime-Switching Tensor Regressions (Non-linearity)

Multiple-equation tensor regression with a shared common latent states + efficient MCMC with back-fitting and random scan.

Chapter 2: Compression (scalability)

Compressed Bayesian tensor regression via generalized tensor random projections: theory (concentration inequality, posterior consistency) + prediction with BMA.

Chapter 3: Time-varying uncertainty

Bayesian tensor regression with stochastic volatility: new augmented SV specifications + SV samplers (MH / AMS).

Chapter 1

Markov Switching Multiple-Equation Tensor Regression

Contributions

Challenges: regime shifts, structural breaks, model misspecification, ...

Contributions

Challenges: regime shifts, structural breaks, model misspecification, ...

A new Bayesian tensor model for multiple-equation regressions that accounts for latent regime changes is proposed.

A We extend the soft tensor linear regression models (Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.

Contributions

Challenges: regime shifts, structural breaks, model misspecification, ...

A new Bayesian tensor model for multiple-equation regressions that accounts for latent regime changes is proposed.

A We extend the soft tensor linear regression models (Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.

B We consider a multiple-equation setting where the tensor coefficients are driven by a common hidden Markov chain process.

Università Ca' Foscari Venezia

Contributions

Challenges: regime shifts, structural breaks, model misspecification, ...

A new Bayesian tensor model for multiple-equation regressions that accounts for latent regime changes is proposed.

- A We extend the soft tensor linear regression models (Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.
- B We consider a multiple-equation setting where the tensor coefficients are driven by a common hidden Markov chain process.
- C A Bayesian inference procedure that based on backfitting and partial random scan Gibbs sampler reduces computational costs and improves scalability.

The Model

Vector-on-Tensor Markov-switching Regression

A system of N equations with time-varying parameters

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

$t = 1, 2, \ldots, T$, where $y_{\ell,t}$, $\ell = 1, \ldots, N$, are scalar response variables

- $\varepsilon_{\ell,t}$, $\ell = 1, \ldots, N$ are i.i.d. from $\mathcal{N}(0, 1)$
- X_t and $B_\ell(s_t)$ are $p_1 \times \cdots \times p_M$ tensor-valued covariate and coefficient, with M denoting the number of tensor modes
- $\langle \cdot, \cdot \rangle$ denotes the inner product for tensors (Kolda and Bader, 2009a)
- The latent process is a K-state Markov chain process

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

Prior specification

Soft PARAFAC (Papadogeorgou et al., 2021)

$$B_{\ell,k} = \sum_{d=1}^{D} B^{(d)}_{\ell,1,k} \odot \cdots \odot B^{(d)}_{\ell,M,k}, \qquad \odot = \text{Hadamard product.}$$

$B^{(d)}_{\ell,m,k}$ are tensors of the same dimension as $B_{\ell,k}$.

Prior specification

Soft PARAFAC (Papadogeorgou et al., 2021)

$$B_{\ell,k} = \sum_{d=1}^{D} B_{\ell,1,k}^{(d)} \odot \cdots \odot B_{\ell,M,k}^{(d)}, \qquad \odot = \text{Hadamard product.}$$

$B_{\ell,m,k}^{(d)}$ are tensors of the same dimension as $B_{\ell,k}$.

Slice-wise prior

Let $B_{\ell,m,\tilde{j}_m,k}^{(d)}$ be the j_m-th slice of $B_{\ell,m,k}^{(d)}$ along the m-th mode, $j_m = 1, \ldots, p_m$, then for each mode m and slice j_m:

$$\operatorname{vec}\left(B_{\ell,m,\tilde{j}_m,k}^{(d)}\right) \sim \mathcal{N}_{q_m}\left(\gamma_{\ell,m,j_m,k}^{(d)} 1, \phi_{\ell,m,k}^{(d)} I_{q_m}\right), \quad q_m = \prod_{m' \neq m} p_{m'}.$$

with a global–component–mode shrinkage scale

$$\phi_{\ell,m,k}^{(d)} = \tau_{\ell,k}\, \zeta_{\ell,k}^{(d)}\, \kappa_{\ell,m,k}.$$

Università Ca' Foscari Venezia

Prior specification

Second stage

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

where $W^{(d)}_{\ell,m,k} = \mathrm{diag}(w^{(d)}_{\ell,m,1,k}, \ldots, w^{(d)}_{\ell,m,j_m,k}, \ldots, w^{(d)}_{\ell,m,p_m,k})$. This random scale specification allows for shrinkage at different levels.

Third stage

At the third stage of the prior, we borrow from Guhaniyogi et al. (2017) and specify the scale prior distributions to induce shrinkage across components and modes:

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

HMM: transition probabilities

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

Posterior Approximation

Sampling strategy

Data augmentation

The joint posterior $p(\boldsymbol{\theta}|\mathrm{y},\mathrm{X})$ is not tractable, we follow a data augmentation strategy and introduce the joint posterior $p(\boldsymbol{\theta},\mathrm{s}|\mathrm{y},\mathrm{X})$ where:
$\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$, $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k,\ \gamma_k,\ \boldsymbol{\mu}_k,\ \boldsymbol{\sigma}^2_k,\ \mathrm{p}_k,\ \tau_k,\ \boldsymbol{\zeta}_k,\ \boldsymbol{w}_k,\ \boldsymbol{\lambda}_k,\ \boldsymbol{\kappa}^2_k)$,
$\mathrm{y} = (\mathrm{y}_1,\ldots,\mathrm{y}_T)$, $\mathrm{X} = (X_1,\ldots,X_T)$, $\boldsymbol{s} = (s_1,\ldots,s_T)'$

Sampling strategy

At every iteration

1. **Random scan**: randomly select a subset of component indices $\{d_1,\ldots,d_{D^*}\}$ of fixed size D^* from the set $\{1,2,\ldots,D\}$, where $D^* < D$ and a subset of mode indices $\{m_1,\ldots,m_{M^*}\}$ of fixed size M^*, where $M^* < M$ from the set $\{1,2,\ldots,M\}$.
2. **Backfitting**: sample $\beta^{(d)}_{\ell,m,j_m,k} := \mathrm{vec}\left(B^{(d)}_{\ell,m,\tilde{j}_m,k}\right)$ from $f(\boldsymbol{\beta}^{(d)}_{\ell,m,j_m,k}|\mathrm{y},\boldsymbol{X},\mu_{\ell,k},\boldsymbol{\beta}_{\ell,-j_m,k},\sigma^2_{\ell,k},\tau_{\ell,k},\zeta^{(d)}_{\ell,k},w^{(d)}_{\ell,m,j_m,k},\kappa^2_{\ell,m,k})$ for $d \in \{d_1,\ldots,d_{D^*}\}$ and $m \in \{m_1,\ldots,m_{M^*}\}$ which is a multivariate normal distribution.

Posterior Approximation

MCMC

We propose a MCMC procedure based on Gibbs sampling to sample the unknowns from 3 blocks.

- Block 1: Sampling $\beta^{(d)}_{\ell,m,j_m,k}, \gamma^{(d)}_{\ell,m,j_m,k}, \kappa^2_{\ell,m,k}, \sigma^2_{\ell,k}, \mu_{\ell,k}$ from $p\left(\beta^{(d)}_{\ell,m,j_m,k}, \gamma^{(d)}_{\ell,m,j_m,k}, \kappa^2_{\ell,m,k}, \sigma^2_{\ell,k}, \mu_{\ell,k} \mid \boldsymbol{Y}, \boldsymbol{X}\right)$
- Block 2: Sampling $\zeta^{(d)}_{\ell,k}$ and $\tau_{\ell,k}$ from $p(\zeta^{(d)}_{\ell,k}, \tau_{\ell,k} \mid \boldsymbol{B}_{\ell,k}, \boldsymbol{\gamma}_{\ell,k}, \boldsymbol{w}_{\ell,k})$
- Block 3: Sampling $\lambda^{(d)}_{\ell,m,k}$ and $w^{(d)}_{\ell,m,j_m,k}$ from $p(\lambda^{(d)}_{\ell,m,k}, w^{(d)}_{\ell,m,j_m,k}|\gamma^{(d)}_{\ell,m,j_m,k}, \tau_{\ell,k}, \zeta^{(d)}_{\ell,k})$

For the hidden states, we apply a Forward Filtering Backward Sampling (FFBS) strategy:

- Draw transitional probabilities $(p_{1k}, \ldots, p_{Kk})$ from Dirichlet distribution $p\left(p_{1k}, \ldots, p_{Kk} \mid \boldsymbol{s}\right)$.
- Compute iteratively the vector of smoothed probabilities $\xi_{t|T}$ by using Hamilton Filter, and draw the state vector s_t from a multinomial distribution $\mathcal{M}\left(1, \xi_{t|T}\right)$.
- Identification constraints are imposed during post-processing to solve the label switching problem. $\mu_{\ell,1} < \cdots < \mu_{\ell,K}$ or $\sigma^2_{\ell,1} < \cdots < \sigma^2_{\ell,K}$.

Numerical Results

Simulation

We evaluate the performance of the proposed models, both TR and MSTR, in an extensive simulation study.

- With true matrix-valued coefficients ranging from different ranks and different levels of sparsity.
- With varying values of $D \in \{2, 3, 4, 5, 6, 7\}$ and $K \in \{1, 2, 3\}$.
- We test the robustness of the proposed models to model misspecification by simulating data and fitting the models under differnt values of K.

Numerical Results

Simulation

- With varying values of $D \in \{2, 3, 4, 5, 6, 7\}$ and $K \in \{1, 2, 3\}$.

Table: WAIC-based model comparison for Markov-Switching Tensor Regression. D is the number of components for tensor decomposition, and K is the number of regimes. The model with the best performance is shown in boldface.

	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$	$D = 7$
$K = 2$	2228.04	**2211.74**	2412.63	2613.94	2475.65	2558.18
$K = 3$	5886.07	4095.90	5991.04	4169.02	7977.52	4257.44

Numerical Results
Model misspecification

- We test the robustness of the proposed models to model misspecification by simulating data and fitting the models under diffrent values of K.

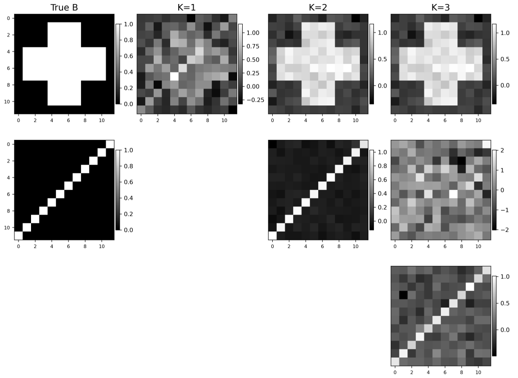


Figure: Parameter recovery comparison of mis-specified models

Numerical Results

Model misspecification

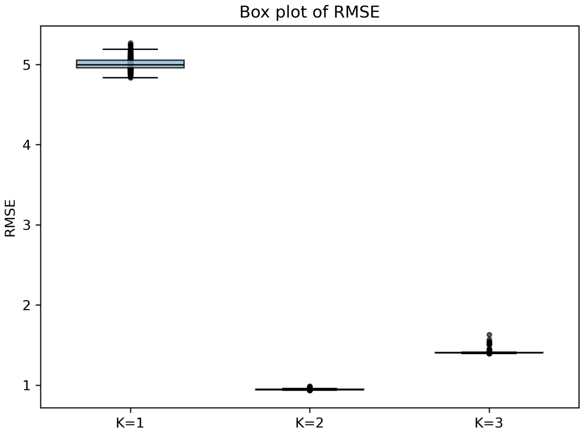


Figure: Box plots of RMSE of the mis-specified models

Oil prices on stock returns

Goals

- We contribute to the debate on the interdependence between financial and oil markets (see, e.g., Xiao and Wang, 2022; Xiao et al., 2023)
- We examine the impact of oil price volatility on the stock market returns (S&P 500) at an aggregate level and on the financial sector, energy sector and other sectors of S&P 500 at the disaggregate level.

Variables

- Oil price volatility is classified into Good Oil Volatility (GV), where the realized volatility is positive, and Bad Oil Volatility (BV), where the realized volatility is negative.
- Other covariates are the Exchange Rate Volatility (ER), TED Spread Volatility (IR) and VIX Index Volatility (VI), following a similar specification as in Xiao and Wang (2022).

Oil prices on stock returns

Model

Specification

- We consider a Mixed Data Sampling (Rodriguez and Puggioni, 2010).
- $y_{\ell,t}$ is the 4-week log-return of market ℓ at time t, $\ell = 1$ (S&P 500), 2 (financial sector), 3 (energy sector) and 4 (other S&P 500 sectors).
- Covariates sampled weekly at the 1st week, 2nd week, 3rd week, 4th week before month t: $t - 1/4, t - 2/4, t - 3/4, t - 4/4$.
- $\mathcal{X}_t \in \mathbb{R}^{5\times4\times4}$: variables $\times$ weeks $\times$ lags.

$$y_{\ell,t} = \mu_\ell(s_t) + \langle \mathcal{B}_\ell(s_t), \mathcal{X}_t \rangle + \sigma_\ell(s_t)\varepsilon_{\ell,t}$$

Oil prices on stock returns

Aggregate vs. Disagregate Regime Detection

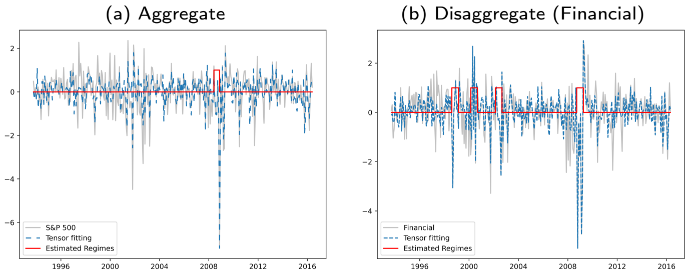


Figure: In-sample fitting on aggregate (left) and disaggregate (right) data, for MSTR (blue dashed line) and estimated hidden states (red solid line). A solid silver line shows actual data.

1997 Asia financial crisis, 2001 9/11 terrorist attack and 2002 corporate scandals and dot-com bubble together with the 2008 global financial crisis.

Concluding Remarks

- A new multiple-equation tensor regression model with Markov switching is proposed.
- An efficient MCMC sampler is proposed based on random scan and back-fitting strategies.
- The tensor regression model is readily to be used with tensor covariates with order 2, 3 or higher.
- Casarin, R., Radu, C., Wang, Q. (2025), Markov Switching Multiple-equation Tensor Regressions, Journal of Multivariate Analysis, 208, 105427

Università Ca' Foscari Venezia

Chapter 2

Compressed Bayesian Tensor Regression

Motivation

- Dimensionality reduction has been a key area of interests in learning from high-dimensional data, to cope with $n << p$ problems.

Motivation

- Dimensionality reduction has been a key area of interests in learning from high-dimensional data, to cope with $n << p$ problems.
- Traditional techniques (e.g., PCA, LDA, SDR) are computationaly intensive and not designed for tensor-valued data (Dasgupta, 2013).

Motivation

- Dimensionality reduction has been a key area of interests in learning from high-dimensional data, to cope with $n << p$ problems.
- Traditional techniques (e.g., PCA, LDA, SDR) are computationaly intensive and not designed for tensor-valued data (Dasgupta, 2013).
- Random projection is computationally efficient and has been successfully applied in many fields, however, its application in tensor-valued data is still under-explored in literature.

Università Ca' Foscari Venezia

Contributions

Random projection for tensor-valued data:

- We propose a generalized tensor random projection: where modes can be projected separately, jointly or preserved.

Contributions

Random projection for tensor-valued data:

- We propose a generalized tensor random projection: where modes can be projected separately, jointly or preserved.
- Concentration inequalities for the proposed tensor projection.

Contributions

Random projection for tensor-valued data:

- We propose a generalized tensor random projection: where modes can be projected separately, jointly or preserved.
- Concentration inequalities for the proposed tensor projection.

Modeling and Inference:

- Apply RP to Bayesian tensor regressions (Guhaniyogi et al., 2017; Guhaniyogi, 2020; Billio et al., 2022; Luo and Griffin, 2025; Casarin et al., 2025a). We consider scalar–on–tensor linear regressions.
- Provide Markov chain Monte Carlo procedures for posterior approximation under alternative prior specifications.
- Provide posterior consistency results built on general theory of Jiang (2007).

Università Ca' Foscari Venezia

A Compressed Bayesian Tensor Regression (CBTR)

Tensor regression

$$y_j = \mu + \langle \mathcal{B}, \mathtt{GTRP}(\mathcal{X}_j) \rangle + \sigma\varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0,1) \tag{2}$$

where $j = 1, \dots, n$, $\mathcal{B} \in \mathbb{R}^{q_1 \times \cdots \times q_M}$ is the coefficient tensor, $\mathcal{X}_j \in \mathbb{R}^{p_1 \times \cdots \times p_N}$ is the covariate tensor for the jth observation.

Università Ca' Foscari Venezia

A Compressed Bayesian Tensor Regression (CBTR)

Tensor regression

$$y_j = \mu + \langle \mathcal{B}, \mathtt{GTRP}(\mathcal{X}_j) \rangle + \sigma \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0,1) \tag{2}$$

where $j = 1, \dots, n$, $\mathcal{B} \in \mathbb{R}^{q_1 \times \dots \times q_M}$ is the coefficient tensor, $\mathcal{X}_j \in \mathbb{R}^{p_1 \times \dots \times p_N}$ is the covariate tensor for the jth observation.

Generalized Tensor Random Projection ($\mathtt{GTRP}$): $\mathbb{R}^{p_1 \times \dots \times p_N} \to \mathbb{R}^{q_1 \times \dots \times q_M}$

$$\mathtt{GTRP}(\mathcal{X}_j) := \mathcal{X}_j \times_1 H_1 \times_2 \dots \times_R H_R \times_{R+1:N} \mathcal{H}_{R+1:N}, \tag{3}$$

- with $R < M \leq N$, where $\mathcal{X} \in \mathbb{R}^{p_1 \times \dots \times p_N}$ is a covariate tensor
- $\times_n$ and $\times_{n:m}$ denote the n-mode and the n-to-m mode products (Kolda and Bader, 2009b)
- $H_m \in \mathbb{R}^{q_m \times p_m}$, $m = 1, \dots, R$ are random projection matrices.
- $\mathcal{H} \in \mathbb{R}^{q_{R+1} \times \dots \times q_M \times p_{R+1} \times \dots \times p_N}$ is a M-mode random projection tensor.

Concentration inequalities

Theorem (JL inequality for mode-wise random projection)

Let $\mathbb{X}$ be an arbitrary set of n order N tensors in $\mathbb{R}^{p_1 \times \cdots \times p_N}$. Define $\mathtt{GTRP}(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 \ldots \times_N H_N$. Define the multilinear mapping $f(\mathcal{X})$: $\mathbb{R}^{p_1 \times \cdots \times p_N} \to \mathbb{R}^{q_1 \times \cdots \times q_N}$.

Given $\epsilon, \beta > 0$ and a sequence of positive integers q_j $j = 1, \ldots, N$ such that $q(N) \geq q_0$ with

$$q_0 = \frac{4 + 2\beta}{\frac{\epsilon^2}{3^N - 1} - \frac{(3^{N+1} - 2)\epsilon^3}{3(3^N - 1)^3}} \log n, \tag{4}$$

with probability at least $1 - n^{-\beta}$, and for all $\mathcal{U}, \mathcal{V} \in \mathbb{X}$, f satisfies

$$(1 - \epsilon)\|\mathcal{U} - \mathcal{V}\|^2 \leq \|f(\mathcal{U}) - f(\mathcal{V})\|^2 \leq (1 + \epsilon)\|\mathcal{U} - \mathcal{V}\|^2$$

Special case: $N = 1$
$q_0 \approx (4 + 2\beta)(\epsilon^2/2 - \epsilon^3/3)^{-1} \log n$ (Achlioptas, 2003).

Bayesian Tensor Regression - Priors

Specification 1: Independent Gaussian and inverse gamma

$$\mathcal{B} \sim \mathcal{TN}_{p_1,\dots,p_M}(0, \Sigma_1, \dots, \Sigma_M), \quad \mu \sim \mathcal{N}(0, \sigma_\mu^2), \quad \sigma^2 \sim \mathcal{IG}(a, b).$$

Specification 2: PARAFAC hierarchical prior (Guhaniyogi et al., 2017)

Let $\circ$ be the *external product* of vectors, and $\gamma_m^{(d)}$, $m = 1, \dots, M$, $d = 1, \dots, D$ the Parallel Factor (PARAFAC) margins

$$\mathcal{B} = \sum_{d=1}^{D} \gamma_1^{(d)} \circ \dots \circ \gamma_N^{(d)},$$

$$\gamma_m^{(d)} \sim \mathcal{N}_{q_m}(0, \tau\zeta^{(d)} W_m^{(d)}), \; \tau \sim \mathcal{IG}(a_\tau, b_\tau), \; w_{m,j_m}^{(d)} \sim \mathcal{E}xp((\lambda_m^{(d)})^2/2),$$

$$\lambda_m^{(d)} \sim \mathcal{G}a(a_\lambda, b_\lambda), \; (\zeta^{(1)}, \dots, \zeta^{(D)}) \sim \mathcal{D}ir(\alpha, \dots, \alpha)$$

where $W_m^{(d)} = \mathrm{diag}(w_{m,1}^{(d)}, \dots, w_{m,j_m}^{(d)}, \dots, w_{m,q_m}^{(d)})$.

Posterior Consistency

Main Result

Let f_0 denote the true predictive density and f the posterior predictive density under compression. Assume all the covariates are bounded and certain assumptions hold (next slide).

For a sequence ε_n satisfying $0 < \varepsilon_n^2 < 1$ and $n\varepsilon_n^2 \to \infty$,

$$E_{f_0}\pi\left[d(f, f_0) > 4\varepsilon_n \mid (y_j, \mathcal{X}_j)_{j=1}^n\right] \leq 4e^{-n\varepsilon_n^2/2}, \tag{5}$$

Posterior Consistency

Key Conditions

1. Controlled Model Complexity

The compressed dimension grows sublinearly $q_n = o(n)$.

2. Well-Behaved Prior (Gaussian prior)

Eigenvalues of covariance matrices are bounded: $\underline{\lambda}_n \leq \lambda \leq \bar{\lambda}_n$.
Prevents overly diffuse or degenerate priors.

3. Norm Preservation (Gaussian prior)

Random projection approximately preserves $\|\mathcal{X}\|$.

4. Covariate entropy control (PARAFAC prior)

Controls the complexity of the model by bounding the projection norm $\|\mathtt{GTRP}(\mathcal{X}_i)\|$, the PARAFAC component D, and the number of coefficients $D\sum_{m=1}^{M} q_{m,n}$.

5. Appropriate contraction rate ε_n (PARAFAC prior)

The posterior contracts, at a rate slower than n^{-1}, but still converges.

Model averaging

- Bayesian Model Averaging (BMA). Relying on a single random projection: a risky approach, as the projection matrix can be far from optimal. In this chapter, we focus on prediction and propose to use BMA.

Each projection ℓ defines a model $\mathcal{M}_\ell$.

$$f(y_{n+j'} \mid \mathcal{D}) = \sum_{\ell=1}^{L} p_\ell(\mathcal{M}_\ell \mid \mathcal{D})\, f_\ell(y_{n+j'} \mid \mathcal{D}, \mathcal{M}_\ell)$$

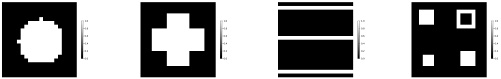
Weights $p_\ell(\mathcal{M}_\ell \mid \mathcal{D})$ estimated via reverse logistic regression (Geyer, 1994).

- Predictive stacking (Gailliot et al., 2024). Instead of posterior model probabilities, choose weights to maximize predictive performance. Particularly appropriate when no model is the true DGP.

Numerical Illustration - Settings

True coefficient values, $\mathcal{B}_0$, in $\langle \mathcal{B}_0, \mathcal{X}_j \rangle$ with iid $\mathcal{X}_j$.

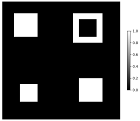
circle (**CI**) cross (**CR**) line (**L**) block (**B**)



- Type of random projection: tensor-wise and mode-wise (1 and 2).
- Covariate tensor dimensions: 20×20 and 60×60 mode-2 tensors.
- Number of observations: from 500 to 2000 at an interval of 500.
- Compression rate, defined as $r = q(M)/p(N)$ with $p(N) = \prod_{m=1}^{N} p_m$, and $q(M) = \prod_{m=1}^{M} q_m$.
- Sparsity coefficient ψ used in generating projection matrices

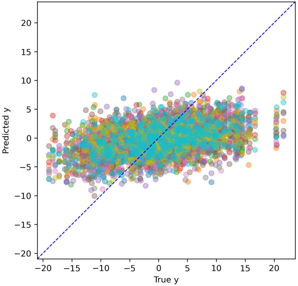
Numerical Illustration - Fitting, 60×60 Block Setting

(a)

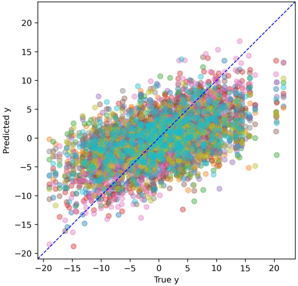


(a)True coefficient, $\mathcal{B}_{\mathbf{0}}$. (b)-(c) Actual data (horizontal axis) against the predicted data (vertical axis) using $L = 10$ independent projection matrices of the same random projection type (colors). Training $n = 1000$, compression rate: $r = 0.36$, sparsity parameter $\psi = 3$.

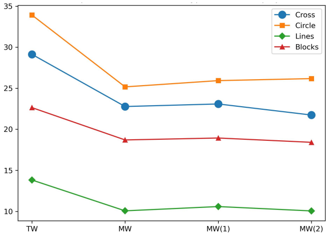
(b) Tensor-wise

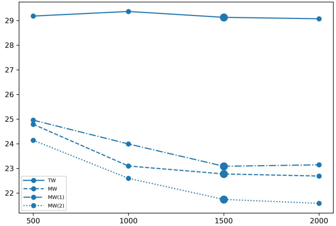


(c) Mode-wise (2)



Numerical Illustration - RMSE, 60×60 Cross Setting





Guidelines for implementation

1. **Type of projection**
   - Prefer mode-preserving projections over tensor-wise.
   - In worst case, mode-wise performs at least as well as tensor-wise.
   - Use exploratory sparsity analysis to decide which modes to compress. For instance, a screen-then-compress strategy, as proposed by Mukhopadhyay and Dunson (2020) or Gailliot et al. (2024), can be adapted for this purpose.
2. **Projection sparsity**
   - Moderate sparsity (e.g. $\psi = 3$) is a good default.
   - Consider more conservative sparsity (e.g. $\psi = 2$) if computation allows.
3. **Model uncertainty**
   - Use Bayesian Model Averaging or Predictive Stacking.
   - Avoid relying on a single projection.

Conclusion

- A new Bayesian tensor regression model with compressed covariates via random projection.
- A new generalized random projection technique to compress tensor structured data.
- Strong theoretical results on concentration properties of random projection and convergency properties of Bayesian inference.
- Few extensions can be considered for future research
  - A pre-screening step to discard predictors with low marginal correlation as proposed by Mukhopadhyay and Dunson (2020) and Gailliot et al. (2024).
  - Bayesian predictive stacking (Gailliot et al., 2024) as an alternative to BMA.
  - Alternative construction of projection tensors (e.g. Kronecker-based, tensor train-based, etc.).
  - Potential applications to data privacy.

Università Ca' Foscari Venezia

Chapter 3

Bayesian Tensor Regression with Stochastic Volatility

Motivation

1. **Volatility is time-varying**
   - Financial returns exhibit volatility clustering.
   - Homoscedastic tensor regression is often unrealistic.
2. **Predictors are high-dimensional and structured**
   - Covariates naturally arise as multi-way arrays (tensors).
   - Structure across modes (time, sector, region) carries information.

Gap in the literature

Scalar-on-tensor regression typically assumes homoscedastic errors, ignoring time-varying volatility.

Contributions

1. **Tensor regression with stochastic volatility**
   - Joint estimation of tensor coefficients and time-varying volatility.
2. **Flexible volatility dynamics**
   - AR-type SV specification.
   - Extension with realized volatility (HAR-style) (Corsi, 2009).
   - Extension with tensor covariates in the volatility equation (Harvey et al., 1994; Koopman et al., 2016).
3. **Efficient Bayesian estimation**
   - Blocked MCMC: tensor block + SV block.
   - MH (Chan and Grant, 2014) and auxiliary mixture sampling (Kim et al., 1998; Sakaria and Griffin, 2017) for latent volatility.

Università Ca' Foscari Venezia

Model: A Bayesian tensor regression with stochastic volatility

$$y_t = \langle \mathcal{B}, \mathcal{X}_t \rangle + e^{h_t/2} \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0,1) \tag{6}$$

1. BTRSV-1

$$h_t = \alpha + \beta(h_{t-1} - \alpha) + \eta_t, \qquad \eta_t \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2) \tag{7}$$

2. BTRSV-2

$$h_t = \alpha + \beta(h_{t-1} - \alpha) + \gamma(h_{t-2} - \alpha) + \eta_t, \qquad \eta_t \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2) \tag{8}$$

Model: A Bayesian tensor regression with stochastic volatility

3. BTRSVRV-1

$$h_t = \alpha + \delta_1 \mathrm{RV}_{t-1} + \delta_2 \mathrm{RV}^{(5)}_{t-1} + \delta_3 \mathrm{RV}^{(22)}_{t-1} + \beta h_{t-1} + \eta_t,\ \eta_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{9}$$

$t = 1, \dots, T$, where RV_{t-1}, $\mathrm{RV}^{(5)}_{t-1}$ and $\mathrm{RV}^{(22)}_{t-1}$ are the daily, weekly and monthly averaged log realized volatility starting from day $t-1$.

4. BTRSVX-1

$$h_t = \alpha + \langle \Gamma, \mathcal{X}_t \rangle + \beta \left(h_{t-1} - \alpha - \langle \Gamma, \mathcal{X}_{t-1} \rangle \right) + \eta_t,\ \eta_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{10}$$

$t = 1, \dots, T$, where $\mathcal{X}_t$ is the same tensor-valued covariates appeared in measurement equation (6), Γ is the tensor-valued coefficients for the latent log volatility.

Posterior approximation - stochastic volatility

We sample α, β and σ^2 from their full conditionals.

Latent log-volatility $\{h_t\}_{t=1}^{T}$ using

- Metroplis-Hastings (Chan and Grant, 2014)
  - Likelihood locally approximated via Taylor expansion.
  - Proposal drawn from approximated posterior.
  - Accept–reject step ensures exact posterior targeting.
  - Flexible and robust for extended SV specifications.

- Auxilairy Mixture Sampler (AMS) (Kim et al., 1998).
  - $\log(\varepsilon_t^2)$ approximated by finite Gaussian mixture.
  - Restores conditional Gaussian state-space structure.
  - Enables efficient Gibbs sampling.
  - Performance sensitive to offset parameter choice (Sakaria and Griffin, 2017).

Simulations
Settings

We carry out simulation studies to demonstrate the validity of the Bayesian inference procedures using AMS and MH for BTRSV-1 and -2.

- For tensor coefficients, we consider a 20×20 cross pattern for the true coefficient $\mathcal{B}_0$ and generate covariate tensors $\mathcal{X}_t$ with i.i.d. standard normal entries.
- For stochastic volatility part, we set

	α	σ	β	γ
SV-1	-1	0.2	0.95	
SV-2			0.5	0.4

Simulations

Results

True Coefficients | BTRSV-1 | BTRSV-2

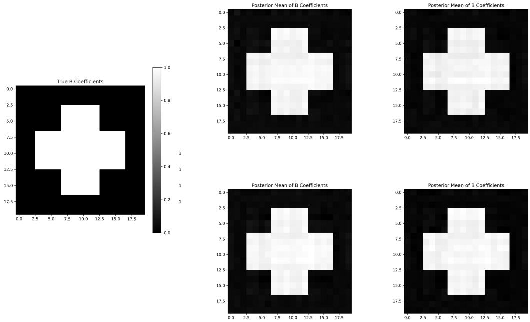


Figure: True tensor-valued coefficients and the estimated posterior median of the coefficients for BTRSV-1 and BTRSV-2 using MH within Gibbs sampler (first row) and auxiliary mixture sampler (second row). The true coefficients are shown in the first column, the estimated coefficients for BTRSV-1 are shown in the second column and the estimated coefficients for BTRSV-2 are shown in the third column.

Simulations

Results

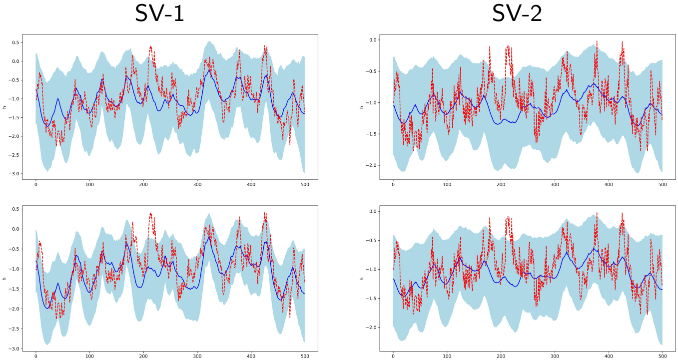


Figure: True and estimated log-volatility for the two stochastic volatility models: BTRSV-1 and BTRSV-2. First row: estimated log-volatility (blue solid line) together with the true log-volatility (red dashed line) and the 95% credible interval (blue shaded area) using MH within Gibbs sampler. Second row: estimated log-volatility (blue solid line) together with the true log-volatility (red dashed line) and the 95% credible interval (blue shaded area) using auxiliary mixture sampler.

Choice of offset parameter c

$\log((y_t - \langle \mathcal{B}, \mathcal{X}_t \rangle)^2 + c)$

Table: Posterior medians of the parameters of SV-1 and SV-2 under different choices of the offset parameter c. The numbers in parentheses are the 95% credible intervals. True values of the parameters are given in the parentheses in the column headers.

SV-1				
$c = 10^{-n}$	$\alpha(-1)$	$\beta(0.95)$	$\sigma_h^2(0.04)$	
$n = 3$	-0.94 (-3.06, 1.40)	0.99 (0.95, 1.00)	0.05 (0.04, 0.05)	
$n = 5$	-0.96 (-3.18, 1.40)	0.98 (0.95, 1.00)	0.05 (0.04, 0.05)	
$n = 7$	-0.98 (-3.23, 1.48)	0.98 (0.95, 1.00)	0.05 (0.04, 0.05)	
$n = 9$	-0.97 (-3.18, 1.38)	0.98 (0.95, 1.00)	0.05 (0.04, 0.05)	
SV-2				
$c = 10^{-n}$	$\alpha(-1)$	$\beta(0.5)$	$\sigma_h^2(0.04)$	$\gamma(0.4)$
$n = 3$	-0.98 (-1.84, -0.10)	0.63 (0.42, 0.86)	0.03 (0.02, 0.03)	0.34 (0.11, 0.56)
$n = 5$	-1.00 (-1.89, -0.05)	0.63 (0.39, 0.85)	0.03 (0.02, 0.03)	0.34 (0.11, 0.59)
$n = 7$	-1.00 (-1.83, -0.13)	0.61 (0.02, 0.86)	0.03 (0.02, 0.03)	0.36 (0.11, 0.96)
$n = 9$	-1.01 (-1.86, -0.17)	0.62 (0.39, 0.86)	0.03 (0.02, 0.03)	0.35 (0.10, 0.58)

Università Ca' Foscari Venezia

Empirical experiment: forecasting market returns

Specification

- y_t is monthly log-return of S&P 500.
- Covariates are sampled daily at 1-day to 22-day before month t: $t - 1/22, t - 2/22, \dots, t - 1$.
- $\mathcal{X}_t \in \mathbb{R}^{7\times 22\times 4}$: variables $\times$ days $\times$ monthly lags.

$$y_t = \langle \mathcal{B}, \mathcal{X}_t \rangle + e^{h_t/2}\varepsilon_t, \quad \varepsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0,1) \tag{11}$$

Empirical experiment: forecasting market returns

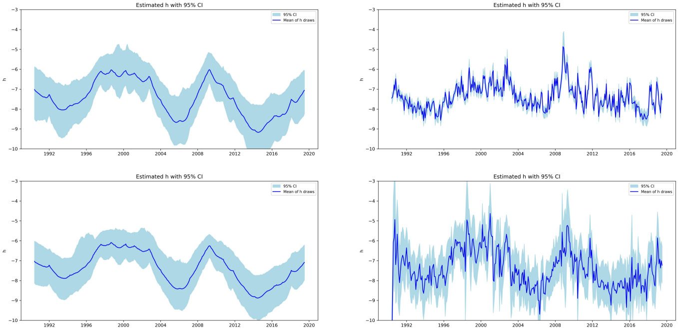


Figure: Estimated log-volatility and their 95% quantiles. Posterior mean of the log-volatility draws (blue line) and 95% credible interval (blue shaded area) for BTRSV-1 and BTRSV-2 (first column), BTRSVRV-1 and BTRSVX-1 (second column)

Empirical experiment: forecasting market returns

We report the Root Mean Square Error (RMSE) and Continuous Ranked Probability Score (CRPS) (Gneiting and Raftery, 2007) to evaluate the in-sample and out-of-sample performance of the seven different models.

Table: Root Mean Squared Error (RMSE) and Continuous Ranked Probability Score (CRPS) for in-sample and out-of-sample forecasts across competing models.

	RMSE				CRPS			
	In-sample		Out-of-sample		In-sample		Out-of-sample	
	MH	AMS	MH	AMS	MH	AMS	MH	AMS
BTRSV-1	0.0267	0.0266	0.0235	0.0235	0.0133	0.0133	0.0138	0.0139
BTRSV-2	0.0270	0.0263	0.0226	0.0239	0.0136	0.0133	0.0132	0.0137
BTRSVRV-1	0.0254	0.0257	0.0218	0.0211	0.0132	0.0133	0.0123	0.0118
BTRSVX-1	0.0260	0.0272	0.0214	0.0242	0.0128	0.0131	0.0128	0.0144
SV-1	0.0417	0.0417	0.0382	0.0380	0.0216	0.0215	0.0211	0.0209
SV-2	0.0416	0.0417	0.0375	0.0380	0.0217	0.0217	0.0205	0.0207
BTR	0.0633		0.0918		0.0397		0.0464	

Università Ca' Foscari Venezia

When to Use BTRSVRV-1 vs BTRSVX-1?

BTRSVRV-1 (SV + Realized Volatility)

- Modest computational cost
- Suitable when realized volatility captures main drivers of risk
- Fast reaction to abrupt market changes

BTRSVX-1 (SV + Tensor Covariates)

- Higher computational burden
- Preferable when covariates contain additional predictive information
- Useful under structural breaks or regime shifts

Practical Guideline

BTRSVRV-1 should be the default choice for volatility modeling unless preliminary analysis indicates strong association between the tensor covariates and volatility dynamics.

Conclusion

- We introduce a unified and flexible Bayesian framework that integrates tensor regression with stochastic volatility modeling.
- We introduce new SV specifications incorporating RV and tensor-valued exogenous variables.
- We provide a scalable and fully Bayesian estimation strategy based on a MH and AMS.
- Empirical study confirms that models that incorporate stochastic volatility exhibit enhanced responsiveness to market conditions and better predictive performance.

Thank You !

I would like to thank my supervisors, committee members, and colleagues for their guidance and support.

Thank you for your attention.

Questions are welcome.

GTRP - Special cases

The GTRP reduces the dimensions of the covariate space, allowing for a smaller number of covariates within each mode, as well as a smaller number of modes. It combines two strategies:

Mode-wise RP

A random projection GTRP-MW is called mode-wise when $\texttt{GTRP-MW}(\mathcal{X}) := \mathcal{X} \times_m H_m$ where $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_N}$ and $H_m \in \mathbb{R}^{q_m \times p_m}$.

$\texttt{GTRP-MW}(\mathcal{X})$ changes the size of mode m from p_m to q_m, keeps the N-mode structure of $\mathcal{X}$.

GTRP - Special cases

The GTRP reduces the dimensions of the covariate space, allowing for a smaller number of covariates within each mode, as well as a smaller number of modes. It combines two strategies:

Mode-wise RP

A random projection GTRP-MW is called mode-wise when $\text{GTRP-MW}(\mathcal{X}) := \mathcal{X} \times_m H_m$ where $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_N}$ and $H_m \in \mathbb{R}^{q_m \times p_m}$.

$\text{GTRP-MW}(\mathcal{X})$ changes the size of mode m from p_m to q_m, keeps the N-mode structure of $\mathcal{X}$.

Tensor-wise RP

A random projection GTRP-TW is called tensor-wise when $\text{GTRP-TW}(\mathcal{X}) := \mathcal{X} \times_{n:m} \mathcal{H}$ where $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_N}$ and $\mathcal{H} \in \mathbb{R}^{q_1 \times \cdots \times q_M \times p_n \times \cdots \times p_m}$, $M \leq N$ and $1 \leq n \leq m \leq N$.

$\text{GTRP-TW}(\mathcal{X})$ changes mode number and size.

GTRP - Projection tensor distribution

Distributions used for the projection tensors

- Entries of H_m and $\mathcal{H}_{R+1:N}$ are iid with mean zero and finite fourth moment (Mukhopadhyay and Dunson, 2020).
- Standard normal distribution. Pros: JL-type inequality (Dasgupta and Gupta, 2003). Cons: dense projections are not well-suited for high-dimensional problems.
- Rademacher distribution is used in Rakhshan and Rabusseau (2021), to encourage sparsity in the projection tensors.

Our choice

We follow Achlioptas (2003) and Li et al. (2006) and assume the entries are independent discrete random variables:

$$r = \begin{cases} \sqrt{\psi} & \text{with prob. } \frac{1}{2\psi} \\ 0 & \text{with prob. } 1 - \frac{1}{\psi} \\ -\sqrt{\psi} & \text{with prob. } \frac{1}{2\psi} \end{cases} \tag{12}$$

CBTR: Concentration inequalities I

Define $c(N, M) = p(N)/q(M)$, $p(N) = \prod_{m=1}^{N} p_m$, and $q(M) = \prod_{m=1}^{M} q_m$. $\mathtt{GTRP}(\mathcal{X})$ preserves the distances between points in the original sample spaces, uniformly in $p(N)$ and N.

Proposition (A JL inequality for tensor-wise random projection)

Let $\mathbb{X}$ be an arbitrary set of n order N tensors in $\mathbb{R}^{p_1 \times \cdots \times p_N}$. Define $\mathtt{GTRP}(\mathcal{X}) = \mathcal{X} \times_{1:N} \mathcal{H}_{1:N}$ with $\mathcal{H}_{1:N}$ an $N+1$ order random tensor in $\mathbb{R}^{p_1 \times \cdots \times p_N \times q_1}$ with entries from the distribution in (12), and the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N)}\,\mathtt{GTRP}(\mathcal{X})$ from $\mathbb{R}^{p_1 \times \cdots \times p_N}$ to $\mathbb{R}^{q_1}$.
Given $\epsilon, \beta > 0$, and a positive integer $q_1 \geq q_0$ where $q_0 = (4 + 2\beta)(\epsilon^2/2 - \epsilon^3/3)^{-1} \log n$, f satisfies with high probability and for all tensors $\mathcal{U}, \mathcal{V} \in \mathbb{X}$:

$$(1-\epsilon)\|\mathcal{U} - \mathcal{V}\|^2 \leq \|f(\mathcal{U}) - f(\mathcal{V})\|^2 \leq (1+\epsilon)\|\mathcal{U} - \mathcal{V}\|^2$$

Proof follows immediately from Achlioptas (2003).

Posterior Consistency - Result 1/2

Theorem (Gaussian/Inverse Gamma prior)

Let $\mathcal{B} \sim \mathcal{TN}(0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N)$ *a priori and* $\tilde{\lambda}_n$ *and* $\underline{\lambda}_n$ *be the largest and smallest eigenvalues of* $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N$*. In addition, assume that all the covariates are bounded, which means* $|x_{jkl}| < 1$ *and* $\lim_{n\to\infty} \sum_{j=1}^{p_{1,n}} \sum_{k=1}^{p_{2,n}} \sum_{l=1}^{p_{3,n}} |b_{jkl,0}| < K$.

Define $D(R) = 1 + R \sup_{|h|\le R} |a'(h)| \sup_{|h|\le R} |\frac{b'(h)}{a'(h)}|$, $\theta_n = \sqrt{q_n p_n}$. *For a sequence* ε_n *satisfying* $0 < \varepsilon_n^2 < 1$ *and* $n\varepsilon_n^2 \to \infty$, *assume that the assumptions* ***A.1***, ***A.2*** *and* ***A.3*** *hold, then*

$$E_{f_0} \pi \left[d(f, f_0) > 4\varepsilon_n \mid (y_j, \mathcal{X}_j)_{j=1}^n \right] \le 4e^{-n\varepsilon_n^2/2}, \tag{13}$$

where $\pi[\cdot \mid (y_j, \mathcal{X}_j)_{j=1}^n]$ *is the posterior measure.*

Proof: verify the sufficient conditions a), b) and c) in Theorem 4 of Jiang (2007).

Università Ca' Foscari Venezia

Posterior Consistency - Result 2/2

Theorem (Hierarchical PARAFAC prior)

Let $\gamma_m^{(d)} \sim \mathcal{N}_{p_m}(0, \tau\zeta^{(d)} W_m^{(d)})$ *a priori, and further assume* $\lim_{n\to\infty} \sum_{j=1}^{p_{1,n}} \sum_{k=1}^{p_{2,n}} \sum_{l=1}^{p_{3,n}} |b_{jkl,0}| < K$ *and that all covariates are standardized, that is,* $|x_{jkl}| < 1$.

For a sequence ε_n *satisfying* $0 < \varepsilon_n^2 < 1$ *and* $n\varepsilon_n^2 \to \infty$*, assume that the assumptions **A.1**, **A.4** and **A.5** hold then*

$$E_{f_0}\pi\left[d(f, f_0) > 4\varepsilon_n \mid (y_i, \mathcal{X}_j)_{j=1}^n\right] \leq 4e^{-n\varepsilon_n^2/2}, \qquad (14)$$

where $\pi[\cdot \mid (y_j, \mathcal{X}_j)_{j=1}^n]$ *is the posterior measure.*

Posterior Consistency

Key Conditions (Gaussian prior)

1. Controlled Model Complexity

$$\frac{q_n \log(1/\varepsilon_n^2)}{n\varepsilon_n^2} \to 0, \quad \frac{\log(q_n)}{n\varepsilon_n^2} \to 0, \quad \frac{q_n \log D(\theta_n \sqrt{8\bar{\lambda}_n n\varepsilon_n^2})}{n\varepsilon_n^2} \to 0$$

.

The compressed dimension grows sublinearly $q_n = o(n)$.

2. Well-Behaved Prior

Eigenvalues of covariance matrices are bounded:

$$\underline{\lambda}_n \leq \lambda \leq \bar{\lambda}_n.$$

Prevents overly diffuse or degenerate priors.

3. Norm Preservation

$$\frac{\log\left(\|\mathtt{GTRP}(\mathcal{X})\|\right)}{n\varepsilon_n^2} \to 0, \quad \|\mathtt{GTRP}(\mathcal{X})\|^2 > 8\frac{(K^2+1)}{B_1}\frac{\log(q_n)}{n\varepsilon_n^2}, \quad \forall \mathcal{X} = \mathcal{X}_1, \ldots,$$

Random projection approximately preserves $\|\mathcal{X}\|$.

Posterior Consistency

Key conditions (PARAFAC Priors)

1. **Controlled Model Complexity**

2. **Covariate entropy control**

$$(\log(\|\mathtt{GTRP}(\mathcal{X}_i)\|) + \log D)\, D \sum_{m=1}^{M} q_{m,n} < Mn\varepsilon_n^2 C$$

for some positive constant C. Controls the complexity of the model by bounding the projection norm $\|\mathtt{GTRP}(\mathcal{X}_i)\|$, the PARAFAC component D, and the number of coefficients $D\sum_{m=1}^{M} q_{m,n}$.

3. **Appropriate contraction rate ε_n**

$$\varepsilon_n^2 = n^{\delta}, \quad \text{with } b - 1 < \delta < 0, \sum_{m=1}^{M} q_{m,n} = \mathcal{O}(n^b)$$

The posterior contracts, at a rate slower than n^{-1}, but still converges. Also, the number of compressed parameters $q_{m,n}$ must grow sublinearly with n. Prevents overfitting as n grows.

CBTR: posterior approximation

The joint posterior distribution $f(\boldsymbol{\gamma}_m^{(d)}, \zeta^{(d)}, \tau, \lambda_m^{(d)}, w_m^{(d)}, \sigma^2, \mu \mid \boldsymbol{y}, \mathtt{GTRP}(\boldsymbol{X}))$ is not tractable, we approximated it using a Gibbs sampling procedure. At each iteration, we draw from the full conditional distributions of each parameter:

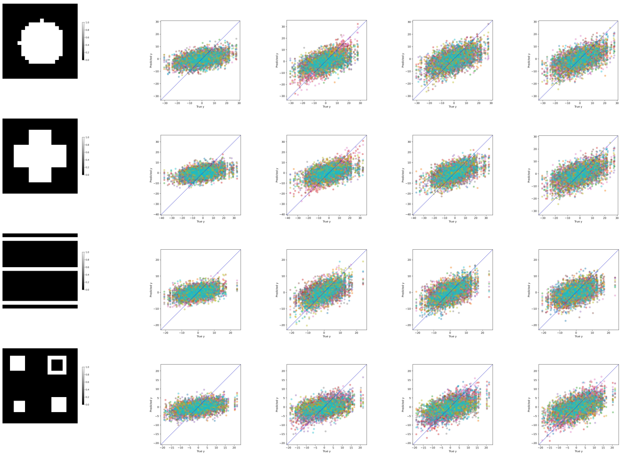
1. Draw $\boldsymbol{\gamma}_m^{(d)}$ from a multivariate normal distribution (back-fitting) $f(\boldsymbol{\gamma}_m^{(d)} \mid \boldsymbol{y}, \mathtt{GTRP}(\boldsymbol{X}), \boldsymbol{\gamma}_{-m}, \tau, \boldsymbol{\zeta}, \boldsymbol{w}, \mu, \sigma^2)$ for $d \in \{1, \dots, D\}, m \in \{1, \dots, M\}$.
2. Draw $\zeta^{(d)}$ from the GIG distribution $f(\zeta^{(d)} \mid \boldsymbol{\gamma}^{(d)}, \tau, \boldsymbol{w}^{(d)})$.
3. Draw τ from the GIG distribution $f(\tau \mid \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{w})$.
4. Draw $\lambda_m^{(d)}$ from $f(\lambda_m^{(d)} \mid \boldsymbol{\gamma}_m^{(d)}, \tau, \zeta^{(d)})$ which is a Gamma distribution.
5. Draw $w_{m,j_m}^{(d)}$ from the GIG distribution $f(w_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)})$.
6. Draw σ^2 from the IG distribution $f(\sigma^2 | \boldsymbol{y}, \mathtt{GTRP}(\boldsymbol{X}), \mu, \boldsymbol{\gamma})$.
7. Draw μ from the Gaussian distribution $f(\mu \mid \boldsymbol{y}, \mathtt{GTRP}(\boldsymbol{X}), \boldsymbol{\gamma}, \sigma^2)$.

CBTR:Numerical Illustration

Fitting, 60 $\times$ 60 All Settings

(a) True Coefficient | (b) Forecast Performance

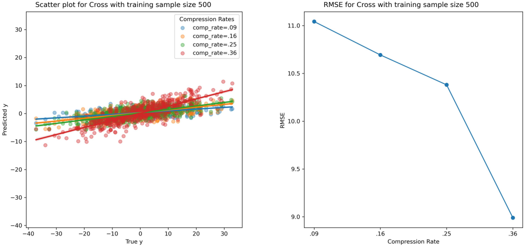
TW | MW | MW(1) | MW (2)



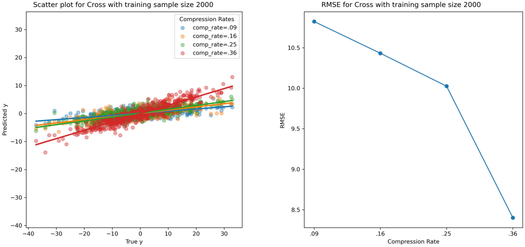
CBTR: Numerical Illustration

Compression rate

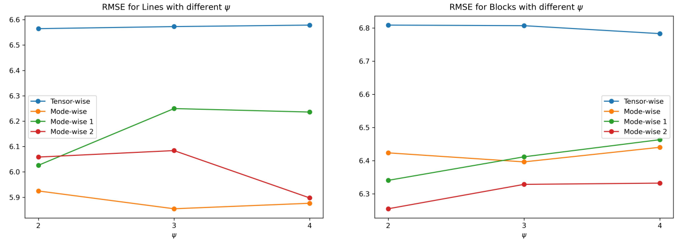
$n = 500$



$n = 2000$



Numerical Illustration - Sparsity



- Projections with different sparsity levels and random projection types: TW (blue), MW (orange), MW(1) (green), and MW(2) (red).
- $m = 500$ test samples, sparsity levels ($\psi \in \{2, 3, 4\}$) (horizontal axis).
- In most scenarios (**CI**, **CR**, and **L**), mode-wise random projection has the lowest RMSE
- V-shape curve for mode-wise suggests a moderate sparsity is preferred.

Università Ca' Foscari Venezia

Clarification of Lemma 2.A.1

Lemma 2.A.1 Let $\mathcal{T} = \boldsymbol{\tau}_1 \otimes \cdots \otimes \boldsymbol{\tau}_N$ be a $q_1 \times \cdots \times q_N$ tensor with $\boldsymbol{\tau}_m \in \mathbb{R}^{q_m}$, $\boldsymbol{\tau}_{m,i_m} \sim \mathcal{N}(0, 1/p_m)$ independent normal. Entries of $\mathcal{T}$ are $\mathcal{T}_{i_1,\dots,i_N} = \boldsymbol{\tau}_{1,i_1} \cdots \boldsymbol{\tau}_{N,i_N}$. Let

$$\mathcal{Q} = \frac{1}{\sqrt{p(N)}} \sum_{j_1=1}^{p_1} \cdots \sum_{j_N=1}^{p_N} H_{1,j_1,:} \otimes \cdots \otimes H_{N,j_N,:}$$

be the $q_1 \times \cdots \times q_N$ tensor obtained by projecting the rescaled $p_1 \times \cdots \times p_N$ unit tensor (a tensor of ones normalized to have unit Frobenius norm) with random matrices H_m defined in Theorem 2.2.1. The tensor entries $Q_{i_1,\dots,i_N}(\mathcal{A})$ of the tensor $\mathcal{Q}(\mathcal{A})$ satisfy the following properties

i. $\mathbb{E}(\mathcal{Q}_{i_1,\dots,i_N}(\mathcal{A})^{2k}) \leq \mathbb{E}(\mathcal{Q}^{2k}_{i_1,\dots,i_N})$

ii. $\mathbb{E}(\mathcal{Q}^{2k}_{i_1,\dots,i_N}) \leq \mathbb{E}(\mathcal{T}^{2k}_{i_1,\dots,i_N})$

Bayesian Inference: Tensor regression

Priors - first stage

We assume that the margins from the PARAFAC decomposition are independent and follow multivariate normal distributions

$$\gamma_m^{(d)} \sim \mathcal{N}_{q_m}(0, \tau\zeta^{(d)} W_m^{(d)}), \qquad m = 1, \dots, M, \quad d = 1, \dots, D \tag{15}$$

Università Ca' Foscari Venezia

Bayesian Inference: Tensor regression

Priors - first stage

We assume that the margins from the PARAFAC decomposition are independent and follow multivariate normal distributions

$$\gamma_m^{(d)} \sim \mathcal{N}_{q_m}(0, \tau\zeta^{(d)} W_m^{(d)}), \qquad m = 1, \ldots, M, \quad d = 1, \ldots, D \quad (15)$$

- τ: global shrinkage parameters.
- $\zeta^{(d)}$: component specific shrinkage parameter, allow a subset of the D factors to contribute more while leaving the values of other components close to zero.
- $W_m^{(d)} = \mathrm{diag}(w_{m,1}^{(d)}, \ldots, w_{m,j_m}^{(d)}, \ldots, w_{m,q_m}^{(d)})$: element specific shrinkage parameter.

Università Ca' Foscari Venezia

Bayesian Inference: Tensor regression

Priors - second stage

We modify the priors from Guhaniyogi and Dunson (2015) and further assume the following prior distributions for the scales:

$$\tau \sim \mathcal{IG}(a_\tau, b_\tau), \quad w_{m,j_m}^{(d)} \sim \mathcal{E}xp((\lambda_m^{(d)})^2/2) \tag{16}$$

$$\lambda_m^{(d)} \sim \mathcal{G}a(a_\lambda, b_\lambda), \quad (\zeta^{(1)}, \dots, \zeta^{(D)}) \sim \mathcal{D}ir(\alpha, \dots, \alpha) \tag{17}$$

Bayesian LASSO: the priors on $w_{m,j_m}^{(d)}$ and $\lambda_m^{(d)}$ lead to Bayesian LASSO type penalty on $\gamma_m^{(d)}$: $\gamma_{m,j_m}^{(d)} \sim \mathcal{DE}\left(0, \sqrt{\tau\zeta^{(d)}}/\lambda_m^{(d)}\right)$ (Park and Casella, 2008).

Posterior approximation - First block

We sample the tensor coefficients $\mathcal{B}$ from $f(\mathcal{B} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})$ and the hyperparameters of its hierarchical prior by using a similar strategy as in the Bayesian tensor regression proposed in Casarin et al. (2025b); Papadogeorgou et al. (2021); Guhaniyogi et al. (2017).

Posterior approximation - First block

We sample the tensor coefficients $\mathcal{B}$ from $f(\mathcal{B} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})$ and the hyperparameters of its hierarchical prior by using a similar strategy as in the Bayesian tensor regression proposed in Casarin et al. (2025b); Papadogeorgou et al. (2021); Guhaniyogi et al. (2017).

1. Draw $\boldsymbol{\gamma}_m^{(d)}$ from a multivariate normal distribution $f(\boldsymbol{\gamma}_m^{(d)} \mid \boldsymbol{y}, \mathcal{X}, \boldsymbol{\gamma}_{-m}, \tau, \boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{h})$ for $d \in \{1, \dots, D\}$ and $m \in \{1, \dots, M\}$.
2. Draw $\zeta^{(d)}$ from the GIG distribution $f(\zeta^{(d)} \mid \boldsymbol{\gamma}^{(d)}, \tau, \boldsymbol{w}^{(d)})$.
3. Draw τ from the IG distribution $f(\tau \mid \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{w})$.
4. Draw $\lambda_m^{(d)}$ from a Gamma distribution $f(\lambda_m^{(d)} \mid \boldsymbol{\gamma}_m^{(d)}, \tau, \zeta^{(d)})$.
5. Draw $w_{m,j_m}^{(d)}$ from the GIG distribution $f(w_{m,j_m}^{(d)} \mid \gamma_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)})$.

Università Ca' Foscari Venezia

Bayesian Inference: Stochastic volatility (BTRSV-1)

Priors for h_t

If we stack all the latent equations by t, we obtain the matrix form of log-volatility process:

$$\boldsymbol{h} = H^{-1}(\boldsymbol{b} + \boldsymbol{\eta}), \qquad \boldsymbol{\eta} \sim \mathcal{N}(0, \Omega)$$

where $\boldsymbol{h} = (h_1, \ldots, h_T)^\top$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)^\top$, $\boldsymbol{b} = (\alpha, \alpha(1-\beta), \ldots, \alpha(1-\beta))^\top$ is a $T \times 1$ vector, $\Omega = \mathrm{diag}(\sigma^2/(1-\beta^2), \sigma^2, \ldots, \sigma^2)$ is a $T \times T$ covariance matrix. H is a $T \times T$ banded matrix. Thus,

$$\boldsymbol{h} \sim \mathcal{N}(H^{-1}\boldsymbol{b}, (H^\top \Omega^{-1} H)^{-1}), \tag{18}$$

Bayesian Inference: Stochastic volatility (BTRSV-1)

Priors for h_t

If we stack all the latent equations by t, we obtain the matrix form of log-volatility process:

$$\boldsymbol{h} = H^{-1}(\boldsymbol{b} + \boldsymbol{\eta}), \qquad \boldsymbol{\eta} \sim \mathcal{N}(0, \Omega)$$

where $\boldsymbol{h} = (h_1, \ldots, h_T)^\top$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)^\top$,
$\boldsymbol{b} = (\alpha, \alpha(1-\beta), \ldots, \alpha(1-\beta))^\top$ is a $T \times 1$ vector,
$\Omega = \mathrm{diag}(\sigma^2/(1-\beta^2), \sigma^2, \ldots, \sigma^2)$ is a $T \times T$ covariance matrix. H is a $T \times T$ banded matrix. Thus,

$$\boldsymbol{h} \sim \mathcal{N}(H^{-1}\boldsymbol{b}, (H^\top \Omega^{-1} H)^{-1}), \tag{18}$$

Priors for α, β, σ

We assume the following priors for α, β, σ^2:

$$\alpha \sim \mathcal{N}(\alpha_0, \sigma_\alpha^2), \qquad \beta \sim \mathcal{N}(\beta_0, \sigma_\beta^2)\mathbb{I}(|\beta| < 1), \qquad \sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma), \tag{19}$$

We impose the stationarity condition $|\beta| < 1$ through the prior on β.

Empirical experiment: forecasting market returns

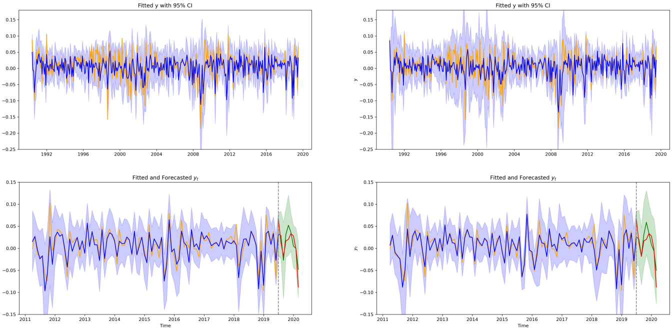


Figure: In-sample and out-of-sample performance. First row: in-sample fitting for BTRSVRV-1 (1st column) and BTRSVX-1 (2nd column). Second row: out-of-sample forecasting for the same models. The observed training response y_t is shown in orange solid line, the posterior medium of estimated response is shown in blue solid line and the 95% credible interval is shown in blue shaded area. The observed test responses are shown in red solid line and the posterior medium of the estimated responses are shown in green solid line with the 95% shown in green shaded area.

Università Ca' Foscari Venezia

References I

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687.

Billio, M., Casarin, R., and Iacopini, M. (2022). Bayesian Markov-Switching Tensor Regression for Time-Varying Networks. *Journal of the American Statistical Association*, pages 1–13.

Casarin, R., Craiu, R. V., and Wang, Q. (2025a). Markov switching multiple-equation tensor regressions. *Journal of Multivariate Analysis*, 208:105427.

Casarin, R., Craiu, R. V., and Wang, Q. (2025b). Markov switching multiple-equation tensor regressions. *Journal of Multivariate Analysis*, 208:105427.

Chan, J. C. C. and Grant, A. (2014). Issues in comparing stochastic volatility models using the deviance information criterion. CAMA Working Paper.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Dasgupta, S. (2013). Experiments with random projection. *arXiv preprint arXiv:1301.3849*.

Università Ca' Foscari Venezia

References II

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Gailliot, S., Guhaniyogi, R., and Peng, R. D. (2024). Data Sketching and Stacking: A Confluence of Two Strategies for Predictive Inference in Gaussian Process Regressions with High-Dimensional Features. *arXiv preprint arXiv:2406.18681*.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in markov chain monte carlo. *Technical Report 568*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Guhaniyogi, R. (2020). Bayesian Methods for Tensor Regression. In Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., and Teugels, J. L., editors, *Wiley StatsRef: Statistics Reference Online*, pages 1–18. Wiley, 1 edition.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian Compressed Regression. *Journal of the American Statistical Association*, 110(512):1500–1514.

References III

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

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *The Annals of Statistics*, 35(4):1487–1511.

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

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

Kolda, T. G. and Bader, B. W. (2009b). Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500.

Koopman, S. J., Lucas, A., and Scharth, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics*, 98(1):97–110.

References IV

Li, P., Hastie, T. J., and Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, Philadelphia PA USA. ACM.

Liu, J., Zhu, C., Long, Z., and Liu, Y. (2021). Tensor Regression. *Foundations and Trends® in Machine Learning*, 14(4):379–565.

Luo, Y. and Griffin, J. E. (2025). Bayesian inference of vector autoregressions with tensor decompositions. *Journal of Business & Economic Statistics*, pages 1–29.

Mukhopadhyay, M. and Dunson, D. B. (2020). Targeted Random Projection for Prediction From High-Dimensional Features. *Journal of the American Statistical Association*, 115(532):1998–2010.

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

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

Rakhshan, B. T. and Rabusseau, G. (2021). Rademacher random projections with tensor networks. *arXiv preprint arXiv:2110.13970*.

Università Ca' Foscari Venezia

References V

Rodriguez, A. and Puggioni, G. (2010). Mixed frequency models: Bayesian approaches to estimation and prediction. *International Journal of Forecasting*, 26(2):293–311.

Sakaria, D. and Griffin, J. E. (2017). On efficient Bayesian inference for models with stochastic volatility. *Econometrics and Statistics*, 3:23–33.

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

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