Compressed Bayesian Tensor Regression

Qing Wang

PhD in Economics
Ca' Foscari University of Venice
Supervisors: Roberto Casarin, Radu Craiu

Pre-Defense
July 2, 2025

PhD thesis

- Chapter 1: Markov-switching multiple-equation tensor regression
  Casarin R., Craiu R., Wang Q.

- Chapter 2: Compressed Bayesian tensor regression
  Casarin R., Craiu R., Wang Q.

- Chapter 3: Bayesian tensor regression with stochastic volatility
  Wang Q.

Chapter 1: Markov-switching multiple-equation tensor regression

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

1. We extend the tensor linear regression models (Guhaniyogi et al., 2017; Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.

Chapter 1: Markov-switching multiple-equation tensor regression

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

1. We extend the tensor linear regression models (Guhaniyogi et al., 2017; Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.
2. We consider a multi-equation setting with possibly different response variables across equations.

Chapter 1: Markov-switching multiple-equation tensor regression

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

1. We extend the tensor linear regression models (Guhaniyogi et al., 2017; Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.
2. We consider a multi-equation setting with possibly different response variables across equations.
3. A low-rank representation of the coefficient tensor and hierarchical prior distribution are proposed to introduce shrinkage effects to overcome overparametrization.

Chapter 1: Markov-switching multiple-equation tensor regression

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

1. We extend the tensor linear regression models (Guhaniyogi et al., 2017; Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.
2. We consider a multi-equation setting with possibly different response variables across equations.
3. A low-rank representation of the coefficient tensor and hierarchical prior distribution are proposed to introduce shrinkage effects to overcome overparametrization.
4. An efficient MCMC sampler is proposed based on back-fitting (Härdle and Hall, 1993) and random scan (Łatuszyński et al., 2013; Yang et al., 2019) strategies.

Chapter 1: Markov-switching multiple-equation tensor regression

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

1. We extend the tensor linear regression models (Guhaniyogi et al., 2017; Papadogeorgou et al., 2021) to an HMM (or MS) framework to accommodate structural breaks.
2. We consider a multi-equation setting with possibly different response variables across equations.
3. A low-rank representation of the coefficient tensor and hierarchical prior distribution are proposed to introduce shrinkage effects to overcome overparametrization.
4. An efficient MCMC sampler is proposed based on back-fitting (Härdle and Hall, 1993) and random scan (Łatuszyński et al., 2013; Yang et al., 2019) strategies.

- Casarin, R., Radu, C., Wang, Q. (2025), Markov Switching Multiple-equation Tensor Regressions, Journal of Multivariate Analysis, 208, 105427
- Casarin, R., Craiu, R., Wang, Q. (2025). Markov Switching Tensor Regressions. In: Aneiros, G., Bongiorno, E.G., Goia, A., Hušková, M. (eds) New Trends in Functional Statistics and Related Fields. IWFOS 2025. Contributions to Statistics. Springer, Cham.

Chapter 3: Bayesian tensor regression with stochastic volatility

1. Introduce a novel Bayesian tensor regression model where the residual variances evolve according to a stochastic volatility (SV) process.
2. Allow for multi-way predictors (e.g., time × asset × feature) and incorporate SV to capture heteroskedasticity common in financial and macroeconomic data.
3. Propose a tailored MCMC sampler for the high-dimensional tensor-SV model that improves mixing and convergence.
4. Compare the performances of different competing SV models in predicting realized volatility on S&P 500.

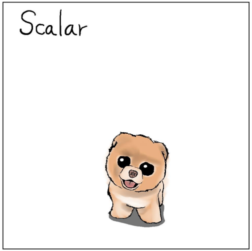
Background: A primer on tensor

Tensors $\Leftrightarrow$ Multi-dimensional array

Background: A primer on tensor

Tensors ⇔ Multi-dimensional array

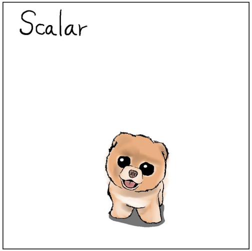
Mode -0



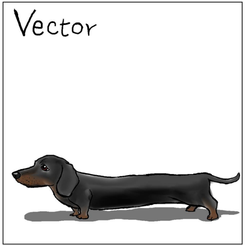
Background: A primer on tensor

Tensors ⇔ Multi-dimensional array

Mode -0



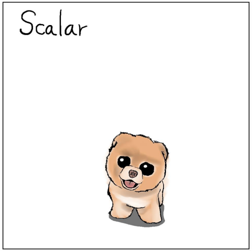
Mode -1



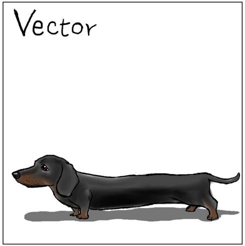
Background: A primer on tensor

Tensors ⇔ Multi-dimensional array

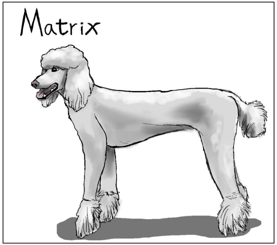
Mode –0



Mode –1



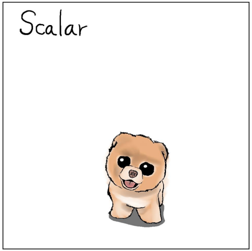
Mode –2



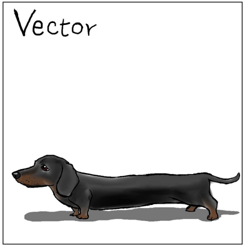
Background: A primer on tensor

Tensors ⇔ Multi-dimensional array

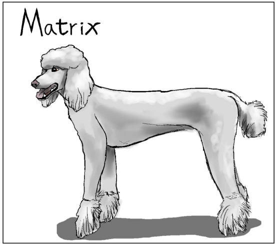
Mode –0



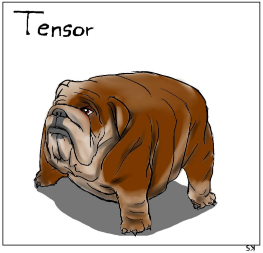
Mode –1



Mode –2

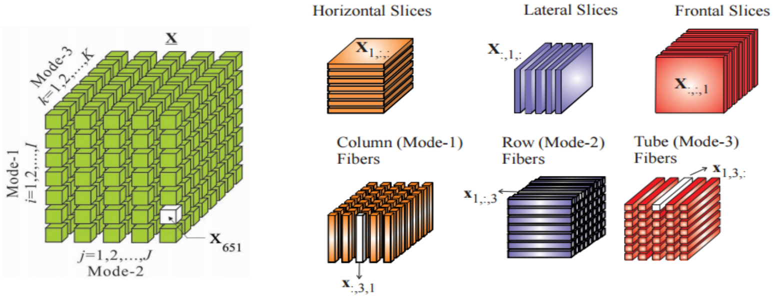


Mode –3



Background: A primer on tensor

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



Background: Tensor regression

Linear regression:

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

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

Background: 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:

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

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

Background: Tensor decomposition

Several **tensor representations/decompositions** available (Tucker, PARAFAC, . . .)

PARAFAC(R) decomposition

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

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

Background: Tensor decomposition

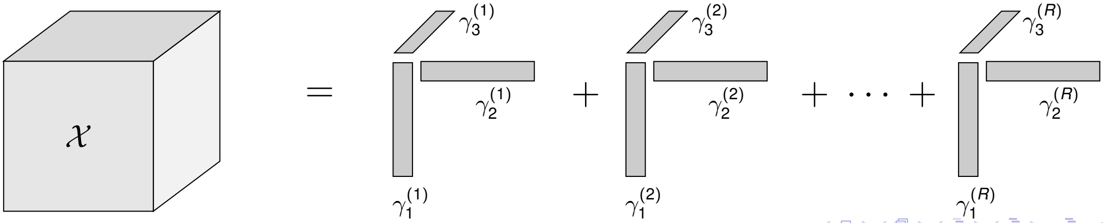
Several **tensor representations/decompositions** available (Tucker, PARAFAC, ...)

PARAFAC(R) decomposition

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

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

Remark: multi-dimensional analogue of **matrix low rank** decomposition.



Background: Tensor multiplication

n-mode product: multiplying a tensor with a matrix

Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and $U \in \mathbb{R}^{J \times I_n}$, the *n*-mode product between $\mathcal{X}$ and U is denoted by $\mathcal{X} \times_n U$, and defined elementwise as

$$(\mathcal{X} \times_n U)_{i_1 \ldots i_{n-1} j \, i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} \mathcal{X}_{i_1 \ldots i_N} u_{j i_n}$$

the result is a tensor of size $I_1 \times \cdots \times I_{n-1} \times \boldsymbol{J} \times I_{n+1} \times \cdots \times I_N$ (Kolda and Bader, 2009).

Background: Tensor multiplication

n-mode product: multiplying a tensor with a matrix

Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and $U \in \mathbb{R}^{J \times I_n}$, the n-mode product between $\mathcal{X}$ and U is denoted by $\mathcal{X} \times_n U$, and defined elementwise as

$$(\mathcal{X} \times_n U)_{i_1 \ldots i_{n-1} j \, i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} \mathcal{X}_{i_1 \ldots i_N} u_{j i_n}$$

the result is a tensor of size $I_1 \times \cdots \times I_{n-1} \times \boldsymbol{J} \times I_{n+1} \times \cdots \times I_N$ (Kolda and Bader, 2009).

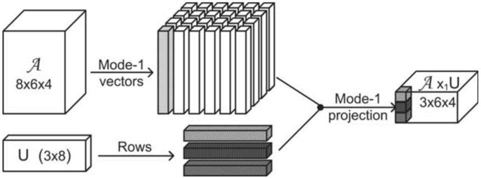


Figure: Visualization of 1− mode product

Background: Tensor multiplication

n−to−*m* mode product (contract product)

Let $\mathcal{X} \in \mathbb{R}^{J_1 \times \ldots \times J_N \times I_1 \times \ldots \times I_M}$ and $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots \times I_M \times K_1 \times \ldots \times K_P}$, the $n-$to$-m$ mode product between $\mathcal{X}$ and $\mathcal{Y}$ is denoted by $\mathcal{X} \times_{N+1:N+M} \mathcal{Y}$, and defined elementwise as

$$(\mathcal{X} \times_{N+1:N+M} \mathcal{Y})_{j_1 \ldots j_N k_1 \ldots k_P} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_M=1}^{I_M} \mathcal{X}_{j_1 \ldots j_n i_1 \ldots i_n} \mathcal{Y}_{i_1 \ldots i_n k_1 \ldots k_p}$$

the result is a tensor of size $J_1 \times \cdots \times J_N \times K_1 \cdots \times K_P$.

Chapter 2: Compressed Bayesian Tensor Regression

Motivation

High dimensional data

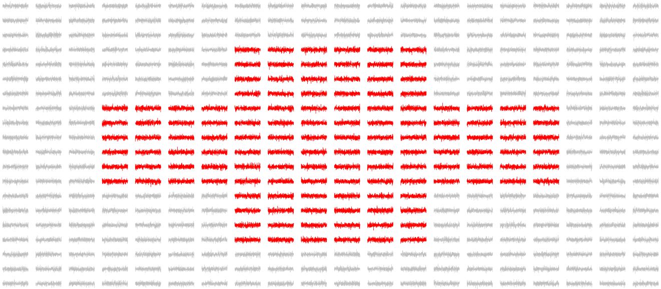
As data grow in volume and complexity, it is increasingly common to record them as high-dimensional arrays or tensors containing massive number of regressors in neuroimaging (Spencer et al., 2022; Guha and Rodriguez, 2021), biostatistics (Clarke et al., 2008) and economics and finance (Billio et al., 2024, 2023). Needs of dimensionality reduction.

Motivation

High dimensional data

As data grow in volume and complexity, it is increasingly common to record them as high-dimensional arrays or tensors containing massive number of regressors in neuroimaging (Spencer et al., 2022; Guha and Rodriguez, 2021), biostatistics (Clarke et al., 2008) and economics and finance (Billio et al., 2024, 2023). Needs of dimensionality reduction.

Raw data (400 covariates)



Motivation

Computational bottleneck and random projections

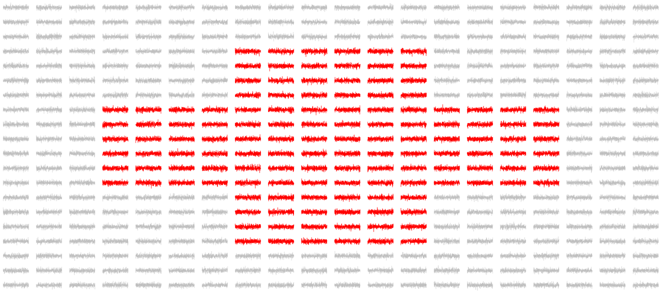
Traditional dimensionality reduction techniques, e.g., PCA, LDA, SDR, despite of their effectiveness are computationally prohibitive when number of regressors is large. Random projection has proven to be effective and computationally efficient (Guhaniyogi and Dunson, 2015; Indyk and Motwani, 1998).

Motivation

Computational bottleneck and random projections

Traditional dimensionality reduction techniques, e.g., PCA, LDA, SDR, despite of their effectiveness are computationally prohibitive when number of regressors is large. Random projection has proven to be effective and computationally efficient (Guhaniyogi and Dunson, 2015; Indyk and Motwani, 1998).

Raw data (400 covariates)



Compressed data (36 covariates)

Contributions

Random projection for tensor

Despite the extensive application of RP, little is studied on applying RP to tensor structured data as well as on their theoretical properties.

1. We propose a generalized tensor random projection (`GTRP`) method that embeds high-dimensional tensor-valued covariates into low-dimensional subspaces with minimal loss of information about the responses.

Contributions

Random projection for tensor

Despite the extensive application of RP, little is studied on applying RP to tensor structured data as well as on their theoretical properties.

1. We propose a generalized tensor random projection (GTRP) method that embeds high-dimensional tensor-valued covariates into low-dimensional subspaces with minimal loss of information about the responses.
2. Strong theoretical support is provided for the concentration properties of the random projection and consistency results of the Bayesian inference.

Contributions

Random projection for tensor

Despite the extensive application of RP, little is studied on applying RP to tensor structured data as well as on their theoretical properties.

1. We propose a generalized tensor random projection (GTRP) method that embeds high-dimensional tensor-valued covariates into low-dimensional subspaces with minimal loss of information about the responses.
2. Strong theoretical support is provided for the concentration properties of the random projection and consistency results of the Bayesian inference.
3. A Bayesian inference framework is provided featuring the use of hierarchical prior distribution and low-rank representation of the parameter.

Random projection

Random projection: a technique of projecting a set of points from a high-dimensional space to a randomly chosen low-dimensional subspace.

Random projection

Random projection: a technique of projecting a set of points from a high-dimensional space to a randomly chosen low-dimensional subspace.

How to project:

1. Let $\boldsymbol{u} = (u_1, \ldots, u_d)^\top$ be a column vector in d-dimensional Euclidean space.
2. Select a k-dimensional subspace represented by $d \times k$ matrix R with $k \in \mathbb{N}$ and $k < d$.
3. The projection will be

$$f(\boldsymbol{u}) = \frac{1}{\sqrt{k}} R^\top \boldsymbol{u} \tag{2}$$

The scaling factor $1/\sqrt{k}$ ensures $\mathbb{E}(\|f(\boldsymbol{u})\|) = \|\boldsymbol{u}\|$.

Random projection

Choosing *R*: Let r_{ij} be the *ij*th entry of *R* and are independently drawn from one of the following distributions satisfying $\mathbb{E}(r_{ij}) = 0, \mathbb{V}(r_{ij}) = 1$:

Random projection

Choosing R: Let r_{ij} be the ijth entry of R and are independently drawn from one of the following distributions satisfying $\mathbb{E}(r_{ij}) = 0, \mathbb{V}(r_{ij}) = 1$:

Gaussian

$$r_{ij} \sim \mathcal{N}(0,1) \quad (3)$$

Random projection

Choosing R: Let r_{ij} be the ijth entry of R and are independently drawn from one of the following distributions satisfying $\mathbb{E}(r_{ij}) = 0, \mathbb{V}(r_{ij}) = 1$:

Gaussian

$$r_{ij} \sim \mathcal{N}(0,1) \tag{3}$$

Rademacher

$$r_{ij} = \begin{cases} +1 & \text{w.p} \quad \frac{1}{2} \\ -1 & \text{w.p} \quad \frac{1}{2} \end{cases} \tag{4}$$

Random projection

Choosing R: Let r_{ij} be the ijth entry of R and are independently drawn from one of the following distributions satisfying $\mathbb{E}(r_{ij}) = 0, \mathbb{V}(r_{ij}) = 1$:

Gaussian

$$r_{ij} \sim \mathcal{N}(0,1) \tag{3}$$

Rademacher

$$r_{ij} = \begin{cases} +1 & \text{w.p} \quad \frac{1}{2} \\ -1 & \text{w.p} \quad \frac{1}{2} \end{cases} \tag{4}$$

Sparse

$$r_{ij} = \sqrt{\psi} \begin{cases} +1 & \text{w.p} \quad \frac{1}{2\psi} \\ 0 & \text{w.p} \quad 1 - \frac{1}{\psi} \\ -1 & \text{w.p} \quad \frac{1}{2\psi} \end{cases} \tag{5}$$

$\psi \in \mathbb{N}$

Random projection

Johnson-Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984)

Given $\varepsilon > 0$ and an integer n, let k be a positive integer such that $k \geq k_0 = O(\varepsilon^{-2} \log n)$, for every set P of n points in $\mathbb{R}^d$ there exists $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $\boldsymbol{u}, \boldsymbol{v} \in P$

$$(1 - \varepsilon) \|\boldsymbol{u} - \boldsymbol{v}\|^2 \leq \|f(\boldsymbol{u}) - f(\boldsymbol{v})\|^2 \leq (1 + \varepsilon) \|\boldsymbol{u} - \boldsymbol{v}\|^2$$

Random projection

Johnson-Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984)

Given $\varepsilon > 0$ and an integer n, let k be a positive integer such that $k \geq k_0 = O(\varepsilon^{-2} \log n)$, for every set P of n points in $\mathbb{R}^d$ there exists $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $\boldsymbol{u}, \boldsymbol{v} \in P$

$$(1 - \varepsilon) \|\boldsymbol{u} - \boldsymbol{v}\|^2 \leq \|f(\boldsymbol{u}) - f(\boldsymbol{v})\|^2 \leq (1 + \varepsilon) \|\boldsymbol{u} - \boldsymbol{v}\|^2$$

Achlioptas (2003)

Let P be an arbitrary set of n points in $\mathbb{R}^d$. Given $\varepsilon, \beta > 0$, for integer $k \geq k_0 = (4 + 2\beta)(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log n$, let R be the $d \times k$ random matrix with entries i.i.d from (4) or (5) and $f : \mathbb{R}^d \to \mathbb{R}^k$ defined in (2). With probability at least $1 - n^{-\beta}$, for all $\boldsymbol{u}, \boldsymbol{v} \in P$

$$(1 - \varepsilon) \|\boldsymbol{u} - \boldsymbol{v}\|^2 \leq \|f(\boldsymbol{u}) - f(\boldsymbol{v})\|^2 \leq (1 + \varepsilon) \|\boldsymbol{u} - \boldsymbol{v}\|^2$$

Literature

Random projection

- Concentration inequalities: Achlioptas (2003); Dasgupta and Gupta (2003); Indyk and Motwani (1998); Frankl and Maehara (1988).
- Sparse RP: Ailon and Chazelle (2009); Indyk and Motwani (1998); Achlioptas (2003); Li et al. (2006); Matoušek (2008).

Literature

Random projection

- Concentration inequalities: Achlioptas (2003); Dasgupta and Gupta (2003); Indyk and Motwani (1998); Frankl and Maehara (1988).
- Sparse RP: Ailon and Chazelle (2009); Indyk and Motwani (1998); Achlioptas (2003); Li et al. (2006); Matoušek (2008).

Tensor random projection

- Tensor to vector: Rakhshan and Rabusseau (2020).
- Tensor to tensor: Li et al. (2021); Shi and Anandkumar (2019).

Literature

Random projection

- Concentration inequalities: Achlioptas (2003); Dasgupta and Gupta (2003); Indyk and Motwani (1998); Frankl and Maehara (1988).
- Sparse RP: Ailon and Chazelle (2009); Indyk and Motwani (1998); Achlioptas (2003); Li et al. (2006); Matoušek (2008).

Tensor random projection

- Tensor to vector: Rakhshan and Rabusseau (2020).
- Tensor to tensor: Li et al. (2021); Shi and Anandkumar (2019).

Posterior consistency

- Bayesian non-parametric: Ghosal et al. (2000); Ghosal and Van Der Vaart (2001)
- Bayesian variable selection: Jiang (2007).
- RP: Guhaniyogi and Dunson (2015); Mukhopadhyay and Dunson (2020).

Literature

Applications

- High-dimensional classification: Chakraborty (2023); Li et al. (2021); Cannings and Samworth (2017).
- Nearest neighbor search: Indyk and Motwani (1998); Datar et al. (2004).
- Image data and face recognition: Bingham and Mannila (2001); Goel et al. (2005).
- Times series and VAR: Farahmand et al. (2017); Koop et al. (2019).
- Data privacy: Li and Li (2023); Gondara and Wang (2020); Anagnostopoulos et al. (2018).

Generalized Tensor Random Projection (GTRP)

A compressed Bayesian tensor regression model

$$y_j = \mu + \langle \mathcal{B}, \mathtt{GTRP}(\mathcal{X}_j) \rangle + \sigma \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0,1) \tag{6}$$

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

Generalized Tensor Random Projection ($\mathtt{GTRP}$)

A compressed Bayesian tensor regression model

$$y_j = \mu + \langle \mathcal{B}, \mathtt{GTRP}(\mathcal{X}_j) \rangle + \sigma \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0,1) \tag{6}$$

where $j = 1, \ldots, n$, $\mathcal{B} \in \mathbb{R}^{q_1 \times \cdots \times q_M}$ 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.

$\mathtt{GTRP}()$: $\mathbb{R}^{p_1 \times \cdots \times p_N} \longrightarrow \mathbb{R}^{q_1 \times \cdots \times q_M}$

$$\mathtt{GTRP}(\mathcal{X}_j) := \mathcal{X}_j \times_1 H_1 \times_2 \ldots \times_R H_R \times_{R+1:N} \mathcal{H}_{R+1:N}, \tag{7}$$

where $\times_n$ and $\times_{n:m}$ denote the n-mode and the n-to-m mode products, $H_m \in \mathbb{R}^{q_m \times p_m}$, $m = 1, \ldots, R$ and $\mathcal{H} \in \mathbb{R}^{q_{R+1} \times \cdots \times q_M \times p_{R+1} \times \cdots \times p_N \times}$ are random projection matrices and M-mode random projection tensor, respectively, with $R \leq M \leq N$.

Generalized Tensor Random Projection

Definition 1 (mode-wise random projection)

A random projection is called mode-wise ($\mathtt{GTRP-M}$) when taking the n-mode product between $\mathcal{X}$ and H_m: $\mathcal{X} \times_m H_m$.

Generalized Tensor Random Projection

Definition 1 (mode-wise random projection)

A random projection is called mode-wise (GTRP−M) *when taking the n-mode product between $\mathcal{X}$ and H_m: $\mathcal{X} \times_m H_m$.*

Example 1 (mode-wise random projection with mode preserving)

Considering a mode-wise random projection for $\mathcal{X} \in \mathbb{R}^{3\times 2}$, $f(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 H_2$, where $H_1 \in \mathbb{R}^{2\times 3}$ random matrix, $H_2 = I_2$, $f(\mathcal{X}) : \mathbb{R}^{3\times 2} \to \mathbb{R}^{2\times 2}$ with the entries,

$$\begin{aligned} f(\mathcal{X})_{i_1,i_2} &= \sum_{j_1=1}^{3}\sum_{j_2=1}^{2} \mathcal{X}_{j_1,j_2} H_{1,i_1,j_1} H_{2,i_2,j_2} \\ &= \sum_{j_1=1}^{3}\sum_{j_2=1}^{2} \mathcal{X}_{j_1,j_2} H_{1,i_1,j_1} \delta(j_2 = i_2) = \sum_{j_1=1}^{3} \mathcal{X}_{j_1,i_2} H_{1,i_1,j_1} \end{aligned}$$

Generalized Tensor Random Projection

Definition 2 (tensor-wise random projection)

A random projection is called tensor-wise ($\mathit{GTRP}-\mathit{T}$) *when taking the n-to-m mode product between* $\mathcal{X}$ *and* $\mathcal{H}$*:* $\mathcal{X} \times_{n:m} \mathcal{H}$.

Generalized Tensor Random Projection

Definition 2 (tensor-wise random projection)

A random projection is called tensor-wise ($\mathtt{GTRP-T}$) *when taking the n-to-m mode product between* $\mathcal{X}$ *and* $\mathcal{H}$*:* $\mathcal{X} \times_{n:m} \mathcal{H}$.

Example 2 (tensor-wise random projection)

Considering a tensor-wise random projection for $\mathcal{X} \in \mathbb{R}^{3\times 2}$, $f(\mathcal{X}) = \mathcal{X} \times_{1:2} \mathcal{H}$, where $\mathcal{H} \in \mathbb{R}^{3\times 2\times 3}$ random tensor, $f(\mathcal{X}) : \mathbb{R}^{3\times 2} \to \mathbb{R}^{3}$ with the entries,

$$f(\mathcal{X})_{i_1} = \sum_{j_1=1}^{3} \sum_{j_2=1}^{2} \mathcal{X}_{j_1,j_2} \mathcal{H}_{j_1,j_2,i_1}$$

Generalized Tensor Random Projection: special cases

(a) If $R = 0$, $M = 1$, GTRP corresponds to the random projection from Nth-order tensor to q_1 dimensional vector: $\mathbb{R}^{p_1 \times \ldots \times p_N} \rightarrow \mathbb{R}^{q_1}$. This setting doesn't exploit the original multiple-mode data structure and it is equivalent to the random projection in Achlioptas (2003) with $d = p_1 \times \ldots \times p_N$ and $k = q_1$ applied to the vectorized tensor.

Generalized Tensor Random Projection: special cases

(a) If $R = 0$, $M = 1$, $\texttt{GTRP}$ corresponds to the random projection from Nth-order tensor to q_1 dimensional vector: $\mathbb{R}^{p_1 \times \ldots \times p_N} \rightarrow \mathbb{R}^{q_1}$. This setting doesn't exploit the original multiple-mode data structure and it is equivalent to the random projection in Achlioptas (2003) with $d = p_1 \times \ldots \times p_N$ and $k = q_1$ applied to the vectorized tensor.

(b) If $R = 0$, $M \geq 1$, only $\texttt{GTRP-T}(\mathcal{X}_j)_{i_1,\ldots,i_M} = \langle \mathcal{X}_j, \mathcal{H}_{i_1,\ldots,i_M,:} \rangle$ is carried out, which returns an M-mode tensor. If $M = N$, the number of modes will be preserved, while only the dimensions along each mode will be reduced. If $M < N$, then not only the dimensions of the tensor will be reduced, but the number of the modes will also be reduced from N to M.

Generalized Tensor Random Projection: special cases

(a) If $R = 0$, $M = 1$, $\texttt{GTRP}$ corresponds to the random projection from Nth-order tensor to q_1 dimensional vector: $\mathbb{R}^{p_1 \times \ldots \times p_N} \to \mathbb{R}^{q_1}$. This setting doesn't exploit the original multiple-mode data structure and it is equivalent to the random projection in Achlioptas (2003) with $d = p_1 \times \ldots \times p_N$ and $k = q_1$ applied to the vectorized tensor.

(b) If $R = 0$, $M \geq 1$, only $\texttt{GTRP-T}(\mathcal{X}_j)_{i_1,\ldots,i_M} = \langle \mathcal{X}_j, \mathcal{H}_{i_1,\ldots,i_M,:} \rangle$ is carried out, which returns an M-mode tensor. If $M = N$, the number of modes will be preserved, while only the dimensions along each mode will be reduced. If $M < N$, then not only the dimensions of the tensor will be reduced, but the number of the modes will also be reduced from N to M.

(c) If $R > 0$, $N = M = R + 1$, only $\texttt{GTRP-M}$ is carried out, where the dimension along each mode is reduced from p_m to q_m, but the number of modes is preserved.

Generalized Tensor Random Projection: special cases

(a) If $R = 0$, $M = 1$, $\texttt{GTRP}$ corresponds to the random projection from Nth-order tensor to q_1 dimensional vector: $\mathbb{R}^{p_1 \times \ldots \times p_N} \to \mathbb{R}^{q_1}$. This setting doesn't exploit the original multiple-mode data structure and it is equivalent to the random projection in Achlioptas (2003) with $d = p_1 \times \ldots \times p_N$ and $k = q_1$ applied to the vectorized tensor.

(b) If $R = 0$, $M \geq 1$, only $\texttt{GTRP-T}(\mathcal{X}_j)_{i_1,\ldots,i_M} = \langle \mathcal{X}_j, \mathcal{H}_{i_1,\ldots,i_M,:} \rangle$ is carried out, which returns an M-mode tensor. If $M = N$, the number of modes will be preserved, while only the dimensions along each mode will be reduced. If $M < N$, then not only the dimensions of the tensor will be reduced, but the number of the modes will also be reduced from N to M.

(c) If $R > 0$, $N = M = R + 1$, only $\texttt{GTRP-M}$ is carried out, where the dimension along each mode is reduced from p_m to q_m, but the number of modes is preserved.

(d) If $R \geq 1$, $M \geq R + 1$, the $\texttt{GTRP}$ involves both mode-wise random projection for the first R modes and tensor-wise random projection for the $(R + 1)$th to Nth modes. Similarly, mode reduction can be performed by choosing $M < N$.

Concentration inequalities I

Corollary 1 (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 $\mathcal{GTRP}(\mathcal{X}) = \mathcal{X} \times_{1:N} \mathcal{H}_{1:N}$ 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 (5), and the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N)}\mathcal{GTRP}(\mathcal{X})$ 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). □

Concentration inequalities II

Theorem 3 (JL inequality for mode-wise random projection)

Let $\mathbb{X}$ be an arbitrary set of n order N tensors in $\mathbb{R}^{p_1 \times \dots \times p_N}$. Define $\mathcal{GTRP}(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 \dots \times_N H_N$, where the entries of $H_m \in \mathbb{R}^{p_m \times q_m}$ for $m = 1, \dots, N$ follows the distribution given in (5)*. Define the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N)}\mathcal{GTRP}(\mathcal{X})$ from $\mathbb{R}^{p_1 \times \dots \times p_N}$ to $\mathbb{R}^{q_1 \times \dots \times q_N}$. Given $\epsilon, \beta > 0$ and a sequence of positive integers q_j $j = 1, \dots, N$ such that $q(N) \geq q_0$ with*

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

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

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

Concentration inequalities II

Theorem 3 (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 $\mathcal{GTRP}(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 \ldots \times_N H_N$, where the entries of $H_m \in \mathbb{R}^{p_m \times q_m}$ for $m = 1, \ldots, N$ follows the distribution given in (5). Define the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N)}\mathcal{GTRP}(\mathcal{X})$ from $\mathbb{R}^{p_1 \times \cdots \times p_N}$ 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

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

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: $N = 1$

$q_0 \approx (4 + 2\beta)(\epsilon^2/2 - \epsilon^3/3)^{-1} \log n$

Bayesian inference: prior assumptions

We specify a hierarchical prior distribution for the regression parameters.

Bayesian inference: prior assumptions

We specify a hierarchical prior distribution for the regression parameters.
In the first stage

$$\gamma_m^{(d)} \sim \mathcal{N}_{q_m}(\mathbf{0}, \tau\zeta^{(d)} W_m^{(d)}), \qquad m = 1, \ldots, M, d = 1, \ldots, D \tag{8}$$

Bayesian inference: prior assumptions

We specify a hierarchical prior distribution for the regression parameters.
In the first stage

$$\gamma_m^{(d)} \sim \mathcal{N}_{q_m}(\mathbf{0}, \tau\zeta^{(d)} W_m^{(d)}), \qquad m = 1, \ldots, M, d = 1, \ldots, D \tag{8}$$

At the second stage, we modify the priors from Guhaniyogi and Dunson (2015) and further assume the following prior distributions for the scales:

$$\tau \sim \mathcal{IG}(a_\tau, b_\tau) \tag{9}$$

$$w_{m,j_m}^{(d)} \sim \mathcal{E}xp((\lambda_m^{(d)})^2/2) \tag{10}$$

$$\lambda_m^{(d)} \sim \mathcal{G}a(a_\lambda, b_\lambda) \tag{11}$$

$$(\zeta^{(1)}, \ldots, \zeta^{(D)}) \sim \mathcal{D}ir(\alpha, \ldots, \alpha) \tag{12}$$

Bayesian inference: convergence properties

Definition 3 (Posterior consistency)

The posterior distribution $\pi_n(\cdot \mid D^{(n)})$ *is said to be weakly (strongly) consistent at* $\theta_0 \in \Theta$ *if* $\pi_n(\theta : d(\theta, \theta_0) > \varepsilon \mid D^{(n)}) \to 0$ *in* $P^{(n)}_{\theta_0}$*-probability (almost surely), as* $n \to \infty$*, for every* $\varepsilon > 0$.

Bayesian inference: convergence properties

Definition 3 (Posterior consistency)

The posterior distribution $\pi_n(\cdot \mid D^{(n)})$ *is said to be weakly (strongly) consistent at* $\theta_0 \in \Theta$ *if* $\pi_n(\theta : d(\theta, \theta_0) > \varepsilon \mid D^{(n)}) \to 0$ *in* $P_{\theta_0}^{(n)}$*-probability (almost surely), as* $n \to \infty$*, for every* $\varepsilon > 0$*.*

Finite-dimensional and parametric models

Doob's theorem (Doob, 1949) and Schwartz's theorem (Schwartz, 1965).

Bayesian inference: convergence properties

Definition 3 (Posterior consistency)

The posterior distribution $\pi_n(\cdot \mid D^{(n)})$ is said to be weakly (strongly) consistent at $\theta_0 \in \Theta$ if $\pi_n(\theta : d(\theta, \theta_0) > \varepsilon \mid D^{(n)}) \to 0$ in $P_{\theta_0}^{(n)}$-probability (almost surely), as $n \to \infty$, for every $\varepsilon > 0$.

Finite-dimensional and parametric models

Doob's theorem (Doob, 1949) and Schwartz's theorem (Schwartz, 1965).

Infinite-dimensional and nonparametric

Contract rate of posterior convergence: *The posterior is said to contract at rate $\varepsilon_n \to 0$ if $\pi_n(f : d(f, f_0) > M_n\varepsilon_n \mid D^{(n)}) \to 0$ in $P_0^{(n)}$-almost surely, for every $M_n \to \infty$ as $n \to \infty$.*

Ghosal et al. (2000) established sufficient conditions to show convergence of posterior measures.

Bayesian inference: convergence properties

High-dimensional with compressed data

- Jiang (2007) established sufficient conditions based on Ghosal et al. (2000) and shows tailored Bayesian variable selection priors lead to near parametric rates in estimating the predictive distribution $f(y \mid x)$.
- Guhaniyogi and Dunson (2015); Mukhopadhyay and Dunson (2020) show that Bayesian regression with compressed data also enjoys similar theoretical guarantees.

Contribution of our paper

- Extension of Guhaniyogi and Dunson (2015); Mukhopadhyay and Dunson (2020) to accommodate tensor-valued covariates.
- Study the consistency under different projection methods and different priors (PARAFAC).

Background: (Jiang, 2007, Theorem 4)

Sufficient conditions

- (a) **Entropy condition**: $\log N(\varepsilon_n, \mathcal{P}_n) \leq n\varepsilon_n^2$ for all large n. Controls the complexity of the model space $\mathcal{P}_n$ by bounding the covering number.
- (b) **Tail mass condition:** $\pi(\mathcal{P}_n^c) \leq e^{-2n\varepsilon_n^2}$ for all large n. Ensures that the prior puts negligible mass outside the model space.
- (c) **Prior concentration condition:** $\pi\left(f : d_t(f, f_0) < \frac{\varepsilon_n^2}{4}\right) \geq e^{-n\varepsilon_n^2/4}$ for all large n. Guarantees that the prior puts enough mass near the true density f_0 (KL neighborhood).

The predictive density is said to contract at rate $\varepsilon_n \to 0$ *if* $\pi_n(f : d(f, f_0) > M_n\varepsilon_n \mid D^{(n)}) \to 0$ *in* $P_0^{(n)}$*-almost surely, for every* $M_n \to \infty$ *as* $n \to \infty$.

Bayesian inference: convergency results I

Theorem 4

Let $\mathcal{B} \sim \mathcal{TN}(\mathbf{0}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_N)$ a priori and $\tilde{\lambda}_n$ and $\underline{\lambda}_n$ be the largest and smallest eigenvalues of $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_N$. Further assume all the covariates are bounded, meaning $|x_{jkl}| < 1$ 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 $D(R) = 1 + Rsup_{|h|\leq R}|a'(h)|sup_{|h|\leq R}|\frac{b'(h)}{a'(h)}|$, $\theta_n = \sqrt{q_n p_n}$. For a sequence ε_n satisfying $0 < \varepsilon_n^2 < 1$, $n\varepsilon_n^2 \to \infty$, assume the following to hold

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

(ii) $\bar{\lambda}_n \leq Bq_n^v, \quad \underline{\lambda}_n \geq B_1 (\log(q_n))^{-1}$

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

for some positive constants B, B_1, v, then

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

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

Bayesian inference: convergency results II

Theorem 5

Let $\gamma_m^{(d)} \sim \mathcal{N}_{p_m}(\mathbf{0}, \tau\zeta^{(d)} W_m^{(d)})$ *a priori, and further assume that all covariates are standardized and bounded, that is,* $|x_{jkl}| < 1$ *and* $\lim_{n\to\infty} \sum_{j=1}^{p_{1,n}} \sum_{k=1}^{p_{2,n}} \sum_{l=1}^{p_{3,n}} |b_{jkl,0}| < K$*. For a sequence* ε_n *satisfying* $0 < \varepsilon_n^2 < 1$, $n\varepsilon_n^2 \to \infty$, *assume that the following hold for some positive constant* C

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

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

then

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

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

Sketch of proof: Setup and Notation

- Tensor predictor: $\mathcal{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_D}$
- Compressed predictor: $\texttt{GTRP}(\mathcal{X}_i)$
- Predictive density: $f(y \mid \langle \mathcal{B}, \texttt{GTRP}(\mathcal{X}_i) \rangle)$
- Hellinger distance: $d(f, f_0) = \int\int (\sqrt{f} - \sqrt{f_0}) \nu_y(dy) \nu_{\mathcal{X}}(d\mathcal{X})$
- Prior: $\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \Sigma_1, \Sigma_2, \Sigma_3)$

Let $\mathcal{P}_n$ be the class of predictive densities induced by $b_{jkl} \in [-b_n, b_n]$, where b_{jkl} is the (jkl)th entry of $\mathcal{B}$. Equivalently: $\mathcal{B} \in [-b_n, b_n]^{q_n}$ where $q_n = \prod_{d=1}^{D} q_{d,n}$.

Sketch of proof: Condition 1: Entropy Bound

We want:

$$\log N(\varepsilon_n, \mathcal{P}_n) \leq n\varepsilon_n^2$$

Sketch:

- Cover $b_{jkl} \in [-b_n, b_n]$ with ℓ_2-balls of radius δ_n
- Lipschitz continuity of GLM ensures:

$$d(f_{\mathcal{B}}, f_{\mathcal{C}}) \leq \|\mathcal{B} - \mathcal{C}\|_2$$

- Choose $\delta_n = \varepsilon_n$ so:

$$\log N(\varepsilon_n, \mathcal{P}_n) \leq q_n \log\left(\frac{b_n}{\varepsilon_n}\right)$$

- Condition is satisfied if:

$$q_n \log\left(\frac{b_n}{\varepsilon_n}\right) \leq n\varepsilon_n^2$$

Sketch of proof: Condition 2: Prior Mass Outside Sieve

We want:

$$\pi(\mathcal{P}_n^c) \leq e^{-2n\varepsilon_n^2}$$

Sketch:

- $\mathcal{P}_n^c = \{\mathcal{B} : \exists jkl, |b_{jkl}| > b_n\}$
- Use Gaussian tail bound:
$$\pi(|b_{jkl}| > b_n) \leq e^{-b_n^2/(2\tilde{\lambda}_n)}$$
- Union bound over q_n dimensions:
$$\pi(\mathcal{P}_n^c) \leq q_n \cdot e^{-b_n^2/(2\tilde{\lambda}_n)}$$
- Choose $b_n = \sqrt{8\tilde{\lambda}_n n\varepsilon_n^2}$ to ensure exponential decay

Sketch of proof: Condition 3: Prior Concentration Near Truth

Goal: Show the prior puts enough mass near the true model f_0 by bounding

$$\pi\left(f : d(f, f_0) < \tfrac{1}{4}\varepsilon_n^2\right) \geq e^{-n\varepsilon_n^2/4}$$

Sketch:

- Let $\mathcal{B}_0$ be the true tensor coefficient and $\langle \mathcal{X}_i, \mathcal{B}_0 \rangle$ the true signal.
- We can show that for all large n: $P\left(|\langle \mathtt{GTRP}(\mathcal{X}_i), \mathcal{B}\rangle - \langle \mathcal{X}_i, \mathcal{B}_0\rangle| < \frac{\varepsilon_n^2}{4\eta}\right) > \exp\left\{-\frac{n\varepsilon_n^2}{4}\right\}$.
- Let $S = \left\{\mathcal{B} : |\langle \mathtt{GTRP}(\mathcal{X}_i), \mathcal{B}\rangle - \langle \mathcal{X}_i, \mathcal{B}_0\rangle| < \frac{\varepsilon_n^2}{4\eta}\right\}$
- $d_{t=1}(f, f_0) = \int\int f_0 \left(\frac{f_0}{f} - 1\right) \nu_y(dy)\nu_{\mathcal{X}}(d\mathcal{X}) = E_{\mathcal{X}}\left[g(u^*)\left(\langle \mathtt{GTRP}(\mathcal{X}_i), \mathcal{B}\rangle - \langle \mathcal{X}_i, \mathcal{B}_0\rangle\right)\right]$.
- Choosing $|g(u^*)| < \eta$ implies that $d_t(f, f_0)$ is a subset of S, hence confirming condition 3.

Bayesian inference: posterior approximation

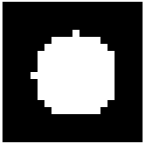
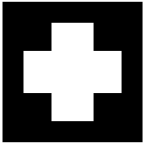
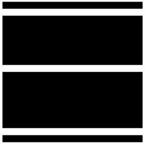
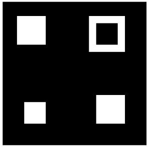
The joint posterior distribution $f(\gamma_m^{(d)}, \zeta^{(d)}, \tau, \lambda_m^{(d)}, w_m^{(d)}, \sigma^2, \mu \mid \boldsymbol{y}, \mathtt{GTRP}(\boldsymbol{X}))$ is not tractable, we approximated it using a Gibbs sampling procedure. The full conditional distributions of the Gibbs sampler are:

Bayesian inference: posterior approximation

The joint posterior distribution $f(\gamma_m^{(d)}, \zeta^{(d)}, \tau, \lambda_m^{(d)}, w_m^{(d)}, \sigma^2, \mu \mid \boldsymbol{y}, \texttt{GTRP}(\boldsymbol{X}))$ is not tractable, we approximated it using a Gibbs sampling procedure. The full conditional distributions of the Gibbs sampler are:

1. Draw $\gamma_m^{(d)}$ from a multivariate normal distribution (back-fitting) $f(\gamma_m^{(d)} \mid \boldsymbol{y}, \texttt{GTRP}(\boldsymbol{X}), \gamma_{-m}, \tau, \zeta, \boldsymbol{w}, \mu, \sigma^2)$ for $d \in \{1, \ldots, D\}, m \in \{1, \ldots, M\}$.
2. Draw $\zeta^{(d)}$ from the GIG distribution $f(\zeta^{(d)} \mid \gamma^{(d)}, \tau, \boldsymbol{w}^{(d)})$.
3. Draw τ from the GIG distribution $f(\tau \mid \gamma, \zeta, \boldsymbol{w})$.
4. Draw $\lambda_m^{(d)}$ from $f(\lambda_m^{(d)} \mid \gamma_m^{(d)}, \tau, \zeta^{(d)})$ 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 $f(\sigma^2 | \boldsymbol{y}, \texttt{GTRP}(\boldsymbol{X}), \mu, \gamma)$.
7. Draw μ from the Gaussian distribution $f(\mu \mid \boldsymbol{y}, \texttt{GTRP}(\boldsymbol{X}), \gamma, \sigma^2)$.

Simulation studies: Settings

	TW	MW
dimensionality	baseline: 20×20 high-dim: 60×60	
tnesor coefficients	 	 
tensor covariates	$x_{ij} \sim \mathcal{N}(0, 1)$	
training sample size	$n \in \{500, 1000, 1500, 2000\}$	
compression rate	$c \in \{.09, .16, .25, .36\}$	
sparsity	$\psi \in \{2, 3, 4\}$	
number of RP	$L = 10$	
number of experiments	7680	

Simulation studies: results I

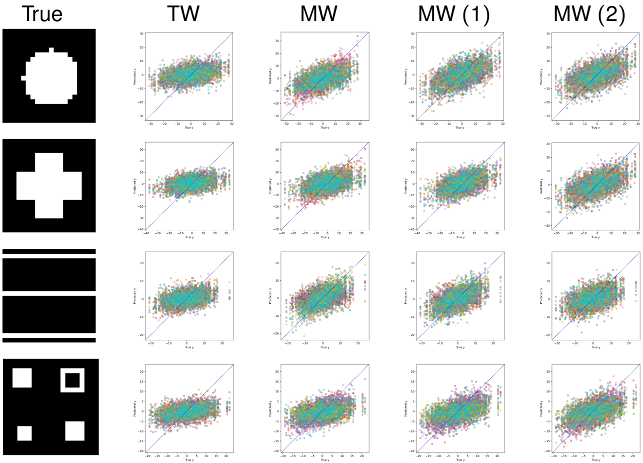


Figure: **Simulation results**: actual data against the predicted for different levels of sparsity (rows) and different types of random projections (columns), using 10 independent projection tensors (colours). For each plot: training sample size: $n = 1000$, compression rate: 0.36, $\psi = 3$.

Simulation studies: results II

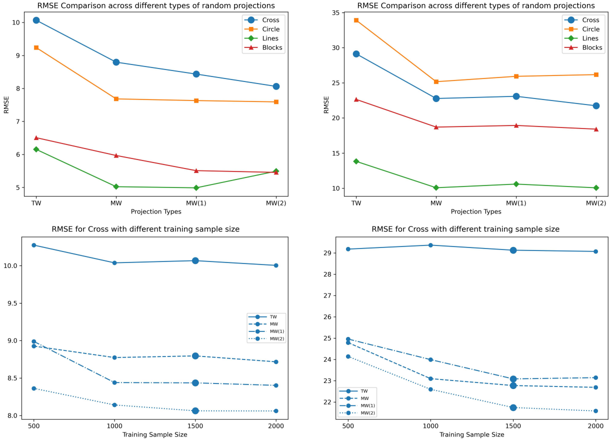


Figure: RMSE comparison across different types of random projection and different configurations in the baseline setting (top) and different sample sizes (bottom) in the 20 × 20 (left) and 60 × 60 dimension case (right). Each estimate is obtained BMA over $L = 10$ independent projection matrices and 500 data points from the validation set.

Simulation studies: results III

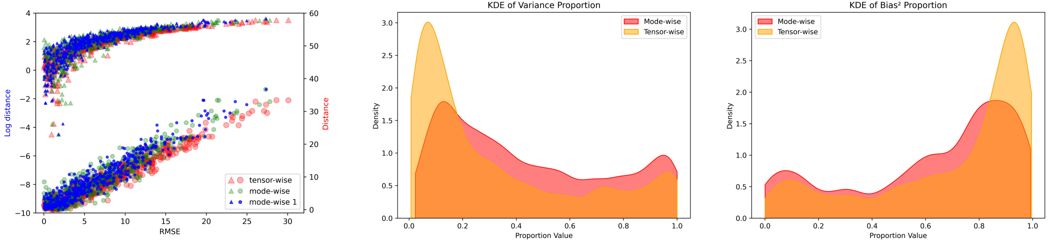


Figure: **Prediction errors**. Left: RMSE vs actual distance (circle, right axis) and log-distance (triangle, left axis) of 500 data points from their mean. Middle: MSE proportion of sample variance. Right: MSE proportion of bias.

Simulation studies: role of ψ

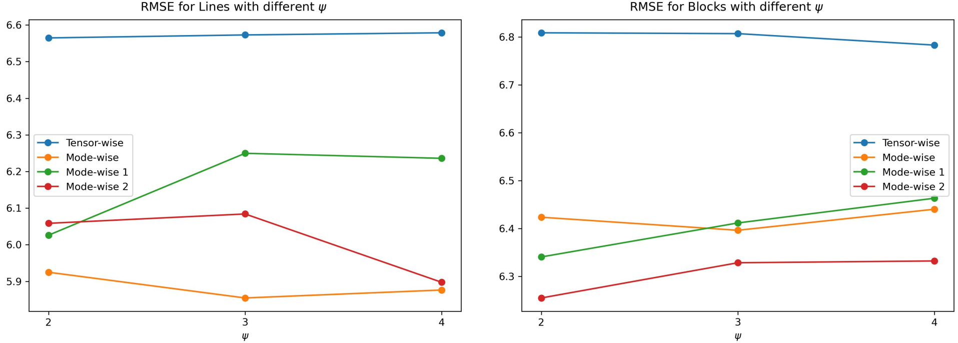


Figure: Effects of different ψ on prediction errors

Simulation studies: role of n and c

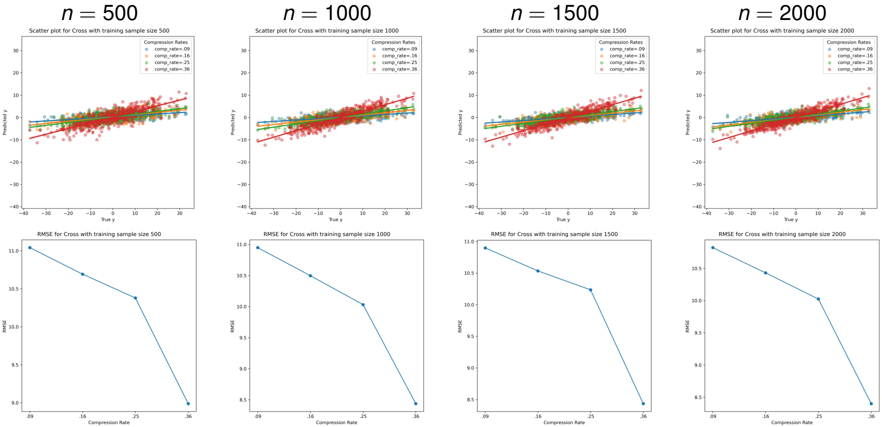


Figure: Comparison of prediction performance for different compression rates $r \in \{0.09, 0.16, 0.25, 0.36\}$ using different training sample size $n \in \{500, 1000, 1500, 2000\}$. Left column: scatter plots for out-of-sample predictions with regression lines for different compression rates. Right column: RMSE of predictions for different compression rates.

Empirical application: macro and financial indicators on stock return

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 compare the performance of different models: BTR, CBTR with different types of random projections (with and without mode preserving).

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), 3-month T-bill rate (TB) and bond spread (BD) following a similar specification as in Xiao and Wang (2022).

Empirical application

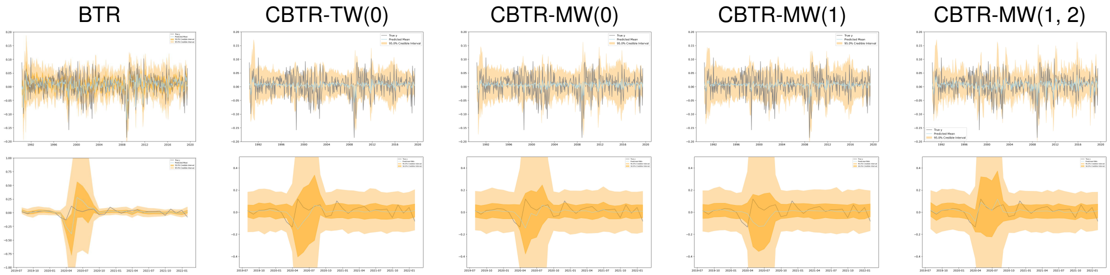
Specification

- Different from Xiao and Wang (2022), we consider a Mixed Data Sampling (Rodriguez and Puggioni, 2010).
- y_t is the montly log-return of market (S&P 500) at time t. Time span: May 1990 to January 2022.
- Covariates sampled daily at the 1st to 22nd day before month t: $t - 1/22, t - 2/22, \dots, t - 22/22$.
- First mode: variables. Second mode: daily data points. Third mode: lagged values.

$$y_t = \mu + \sum_{i_3=1}^{4} \left\langle B_{\tilde{I}(i_3)}, \begin{pmatrix} \mathrm{GV}_{t-\frac{1}{22}-i_3+1} & \mathrm{GV}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{GV}_{t-\frac{21}{22}-i_3+1} & \mathrm{GV}_{t-i_3} \\ \mathrm{BV}_{t-\frac{1}{22}-i_3+1} & \mathrm{BV}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{BV}_{t-\frac{21}{22}-i_3+1} & \mathrm{BV}_{t-i_3} \\ \mathrm{ER}_{t-\frac{1}{22}-i_3+1} & \mathrm{ER}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{ER}_{t-\frac{21}{22}-i_3+1} & \mathrm{ER}_{t-i_3} \\ \mathrm{IR}_{t-\frac{1}{22}-i_3+1} & \mathrm{IR}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{IR}_{t-\frac{21}{22}-i_3+1} & \mathrm{IR}_{t-i_3} \\ \mathrm{VI}_{t-\frac{1}{22}-i_3+1} & \mathrm{VI}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{VI}_{t-\frac{21}{22}-i_3+1} & \mathrm{VI}_{t-i_3} \\ \mathrm{TB}_{t-\frac{1}{22}-i_3+1} & \mathrm{TB}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{TB}_{t-\frac{21}{22}-i_3+1} & \mathrm{TB}_{t-i_3} \\ \mathrm{BD}_{t-\frac{1}{22}-i_3+1} & \mathrm{BD}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{BD}_{t-\frac{21}{22}-i_3+1} & \mathrm{BD}_{t-i_3} \end{pmatrix} \right\rangle + \sigma\varepsilon_t, \tag{13}$$

where $\tilde{I}(i_3) = \{(i_1, i_2, i_3), i_h \in \{1, \dots, p_h\}, \forall h \neq 3\}$ and $B_{\tilde{I}(i_3)}$ denotes the i_3th slice of tensor coefficients B along the third mode.

Empirical application


BTR

CBTR-TW(0)

CBTR-MW(0)

CBTR-MW(1)

CBTR-MW(1, 2)

Figure: Fitting comparison between BTR and CBTR with different random projection methods. First row: in-sample fitting. Second row: out-of-sample prediction. True data are shown in gray solid line, predicted values are shown in blue solid line, light and dark orange colors represent 95% and 50% credible interval, respectively.

Empirical application

Table: RMSE of predictions of BTR and CBTR with different types of random projection methods.

	BTR	CBTR					
		TW	MW	$\text{MW}(1)$	$\text{MW}(1,2)$	$\text{MW}(1,3)$	$\text{MW}(2,3)$
In-sample	0.0338	0.0355	0.0346	0.0356	0.0333	**0.0323**	0.0329
Out-of-sample	0.1148	0.0676	0.0623	0.0723	**0.0383**	0.0600	0.0508

Conclusion

Contributions

- A new random projection technique to compress tensor structured data.
- Strong theoretical results on concentration properties of random projection and convergency properties of Bayesian inference.
- Bayesian compressed tensor regression offers better out-of-sample performance with significant less of computational cost.

References I

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

Ailon, N. and Chazelle, B. (2009). The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322.

Anagnostopoulos, A., Angeletti, F., Arcangeli, F., Schwiegelshohn, C., Vitaletti, A., et al. (2018). Random projection to preserve patient privacy. In *ACM 1st International Workshop on Knowledge Management for Healthcare (KMH2018)*.

Billio, M., Casarin, R., and Iacopini, M. (2024). Bayesian Markov-switching tensor regression for time-varying networks. *Journal of the American Statistical Association*, 119(545):109–121.

Billio, M., Casarin, R., Iacopini, M., and Kaufmann, S. (2023). Bayesian dynamic tensor regression. *Journal of Business & Economic Statistics*, 41(2):429–439.

Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, San Francisco California. ACM.

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):959–1035.

Chakraborty, A. (2023). Efficient Bayesian High-Dimensional Classification via Random Projection with Application to Gene Expression Data. *Journal of Data Science*, pages 1–21.

References II

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature reviews cancer*, 8(1):37–49.

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

Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S. (2004). Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262.

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27.

Farahmand, A.-m., Pourazarm, S., and Nikovski, D. (2017). Random projection filter bank for time series data. *Advances in neural information processing systems*, 30.

Frankl, P. and Maehara, H. (1988). The johnson-lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2).

Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5).

Goel, N., Bebis, G., and Nefian, A. (2005). Face recognition experiments with random projection. page 426, Orlando, Florida, USA.

References III

Gondara, L. and Wang, K. (2020). Differentially private small dataset release using random projections. In *Conference on Uncertainty in Artificial Intelligence*, pages 639–648. PMLR.

Guha, S. and Rodriguez, A. (2021). Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association*, 116(534):581–593.

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

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

Härdle, W. and Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statistica neerlandica*, 47(1):43–57.

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM Symposium on Theory of Computing*, pages 604–613.

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.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. In Beals, R., Beck, A., Bellow, A., and Hajian, A., editors, *Contemporary Mathematics*, volume 26, pages 189–206. American Mathematical Society, Providence, Rhode Island.

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

References IV

Koop, G., Korobilis, D., and Pettenuzzo, D. (2019). Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210(1):135–154.

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.

Li, P., Karim, R., and Maiti, T. (2021). TEC: Tensor Ensemble Classifier for Big Data. Publisher: arXiv Version Number: 1.

Li, P. and Li, X. (2023). Differential privacy with random projections and sign random projections. *arXiv preprint arXiv:2306.01751*.

Matoušek, J. (2008). On variants of the johnson–lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156.

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.

Rakhshan, B. and Rabusseau, G. (2020). Tensorized random projections. In *International Conference on Artificial Intelligence and Statistics*, pages 3306–3316.

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

Schwartz, L. (1965). On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

References V

Shi, Y. and Anandkumar, A. (2019). Higher-order Count Sketch: Dimensionality Reduction That Retains Efficient Tensor Operations. arXiv:1901.11261 [cs, stat].

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

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.

Yang, J., Levi, E., Craiu, R. V., and Rosenthal, J. S. (2019). Adaptive component-wise multiple-try metropolis sampling. *Journal of Computational and Graphical Statistics*, 28(2):276–289.

Łatuszyński, K., Roberts, G. O., and Rosenthal, J. S. (2013). Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability*, 23(1):66 – 98.