Compressed Bayesian Tensor Regression

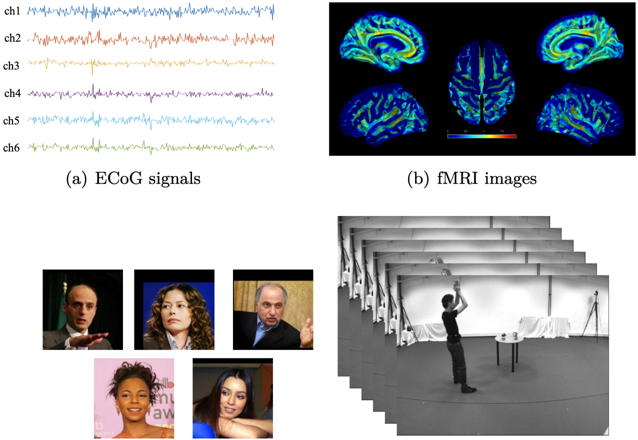
Roberto Casarin[‡], Radu Craiu[†], **Qing Wang**[‡]

[†]University of Toronto
[‡]**Ca' Foscari University of Venice**

Third OCEAN workshop on privacy
March 16, 2026

Background

Tensor Regression



(a) ECoG signals

(b) fMRI images

(c) Facial images

(d) Video sequences

Multi-way data (tensor) (Liu et al., 2021)

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

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, $\mathcal{B}, \mathcal{X}_t \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_M}$.

Motivation

High dimensional data

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

Motivation

High dimensional data

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

Computational bottleneck

Traditional dimensionality reduction techniques, e.g., PCA, LDA, SDR, despite of their effectiveness are computationally prohibitive when number of regressors is large.

Motivation

High dimensional data

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

Computational bottleneck

Traditional dimensionality reduction techniques, e.g., PCA, LDA, SDR, despite of their effectiveness are computationally prohibitive when number of regressors is large.

Random Projection

Random projection is computationally efficient and has been successfully applied in many fields (Johnson and Lindenstrauss, 1984). However, its application in tensor-valued data is still under-explored in literature.

Contribution 1/3 - Projections

Tensorized Random Projections $\mathbb{R}^{p_1 \times \cdots \times p_N} \to \mathbb{R}^{q_1 \times \cdots \times q_M}$

- Random Tensor Train (TT) (Oseledets, 2011) or Canonical Polyadic (CP) low-rank tensor and inner product random tensor and predictor tensor (Rakhshan and Rabusseau, 2020) $\mathbb{R}^{p_1 \times \cdots \times p_N} \to \mathbb{R}^{q}$
- Count Sketch (CS) Charikar et al. (2004) and HCS and *n*-mode product along each mode for 3-mode tensors, preserves the data structure (Shi and Anandkumar, 2019) ($\mathbb{R}^{p_1 \times p_2 \times p_3} \to \mathbb{R}^{q_1 \times q_2 \times q_3}$).

Contribution 1/3 - Projections

Tensorized Random Projections $\mathbb{R}^{p_1 \times \cdots \times p_N} \to \mathbb{R}^{q_1 \times \cdots \times q_M}$

- Random Tensor Train (TT) (Oseledets, 2011) or Canonical Polyadic (CP) low-rank tensor and inner product random tensor and predictor tensor (Rakhshan and Rabusseau, 2020) $\mathbb{R}^{p_1 \times \cdots \times p_N} \to \mathbb{R}^{q}$
- Count Sketch (CS) Charikar et al. (2004) and HCS and *n*-mode product along each mode for 3-mode tensors, preserves the data structure (Shi and Anandkumar, 2019) ($\mathbb{R}^{p_1 \times p_2 \times p_3} \to \mathbb{R}^{q_1 \times q_2 \times q_3}$).

Our contributions

- A generalized tensor random projection: some modes are projected separately, whereas other modes are projected jointly or preserved.
- Concentration inequalities for the proposed tensor projection.

Contribution 2/3 - Modelling

Random Projection (RP) and applications

- Nearest neighbor search Indyk and Motwani (1998); Ailon and Chazelle (2009); Datar et al. (2004)
- High-dimensional classification Chakraborty (2023); Li et al. (2021); Cannings and Samworth (2017)
- Data privacy Li and Li (2023); Gondara and Wang (2020); Anagnostopoulos et al. (2018)
- Inference for large regression models Guhaniyogi and Dunson (2015); Farahmand et al. (2017) and dynamic regressions Koop et al. (2019).

Contribution 2/3 - Modelling

Random Projection (RP) and applications

- Nearest neighbor search Indyk and Motwani (1998); Ailon and Chazelle (2009); Datar et al. (2004)
- High-dimensional classification Chakraborty (2023); Li et al. (2021); Cannings and Samworth (2017)
- Data privacy Li and Li (2023); Gondara and Wang (2020); Anagnostopoulos et al. (2018)
- Inference for large regression models Guhaniyogi and Dunson (2015); Farahmand et al. (2017) and dynamic regressions Koop et al. (2019).

Our contribution

Apply RP to Bayesian tensor regressions (Guhaniyogi et al., 2017; Guhaniyogi, 2020; Billio et al., 2022, 2024; Luo and Griffin, 2025; Casarin et al., 2025). We consider scalar–on–tensor linear regressions.

Contribution 3/3 - Inference

Bayesian Inference and RP

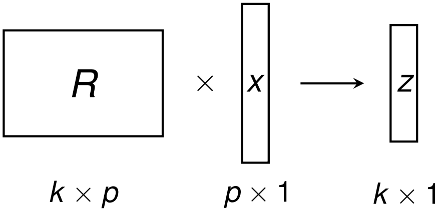
- Bayesian model averaging and posterior consistency for compressed regressions (Guhaniyogi and Dunson, 2015; Mukhopadhyay and Dunson, 2020).
- Data sketching and stacking Gailliot et al. (2024).

Our Contribution

- Provide Markov chain Monte Carlo procedures for posterior approximation under alternative prior specifications.
- Provide posterior consistency results built on general theory of Jiang (2007).

Random Projection

$$z = \frac{1}{\sqrt{k}} R x \tag{1}$$



Key idea:

- The random matrix R compresses a high-dimensional vector $x \in \mathbb{R}^p$
- into a lower-dimensional representation $z \in \mathbb{R}^k$, where $k \ll p$.
- Entries of R are typically sampled as

$$R_{ij} = \sqrt{\psi} \begin{cases} +1 & \text{w.p} \quad \frac{1}{2\psi} \\ 0 & \text{w.p} \quad 1 - \frac{1}{\psi} \\ -1 & \text{w.p} \quad \frac{1}{2\psi} \end{cases}, \qquad \psi \in \mathbb{N} \tag{2}$$

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 $\boldsymbol{u}, \boldsymbol{v} \in P$

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

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 $\boldsymbol{u}, \boldsymbol{v} \in P$

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

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 (2) and $f : \mathbb{R}^d \to \mathbb{R}^k$ defined in (1). With probability at least $1 - n^{-\beta}$, for all $\boldsymbol{u}, \boldsymbol{v} \in P$

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

A Compressed Bayesian Tensor Regression (CBTR)

Tensor regression

$$y_j = \mu + \left\langle \mathcal{B}, \mathtt{GTRP}(\mathcal{X}_j) \right\rangle + \sigma \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0,1) \tag{3}$$

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.

A Compressed Bayesian Tensor Regression (CBTR)

Tensor regression

$$y_j = \mu + \langle \mathcal{B}, \mathtt{GTRP}(\mathcal{X}_j) \rangle + \sigma \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{3}$$

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

$$\mathtt{GTRP}(\mathcal{X}_j) \coloneqq \mathcal{X}_j \times_1 H_1 \times_2 \ldots \times_R H_R \times_{R+1:N} \mathcal{H}_{R+1:N}, \tag{4}$$

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

Concentration inequalities

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

Theorem 1 (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 $\mathtt{GTRP}(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 \ldots \times_N H_N$, where the entries of $H_m \in \mathbb{R}^{p_m \times q_m}$ for $m = 1, \ldots, N$ follows the distribution given in (2)*. Define the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N)}\mathtt{GTRP}(\mathcal{X})$ from $\mathbb{R}^{p_1 \times \cdots \times p_N}$ to $\mathbb{R}^{q_1 \times \cdots \times q_N}$.*

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

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

Concentration inequalities

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

Theorem 1 (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 $\texttt{GTRP}(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 \ldots \times_N H_N$, where the entries of $H_m \in \mathbb{R}^{p_m\times q_m}$ for $m = 1, \ldots, N$ follows the distribution given in (2). Define the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N)}\texttt{GTRP}(\mathcal{X})$ 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 Tensor Regression - Priors

Specification 1: Independent Gaussian and inverse gamma

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

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

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

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

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

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

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:

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:

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

Bayesian Tensor Regression - Model averaging

- Single random projection: a risky approach, as the projection matrix can be far from optimal. We focus on prediction and propose to use Bayesian Model Averaging (BMA).

Bayesian Tensor Regression - Model averaging

- Single random projection: a risky approach, as the projection matrix can be far from optimal. We focus on prediction and propose to use Bayesian Model Averaging (BMA).

BMA predictive density

Let $\mathcal{M}_\ell$ be a model based on $\texttt{GTRP}^{(\ell)}(\cdot)$ with predictive density $f_\ell(\cdot \mid \texttt{GTRP}^{(\ell)}(\mathcal{X}_{n+j'}), \mathcal{D}, \mathcal{M}_\ell)$, $\mathcal{D} = \{(y_j, \texttt{GTRP}(\mathcal{X}_j)), j = 1, \ldots, n\}$, $\boldsymbol{\theta}^{(\ell)} = (\mu^{(\ell)}, \mathcal{B}^{(\ell)}, \sigma^{2^{(\ell)}})$. The BMA predictive density is

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

$j' = 1, \ldots, m$ with m the validation set size.

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

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 assumpstions

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)

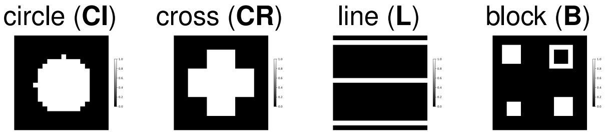
Controls the complexity of the model by bounding the projection norm $\|\mathtt{GTRP}(\mathcal{X}_i)\|$, the PARAFAC component D, and the number of coefficients $D\sum_{m=1}^{M} q_{m,n}$.

5. Appropriate contraction rate ε_n (PARAFAC prior)

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

Numerical Illustration - Settings

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

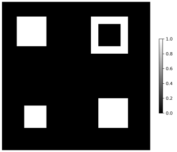


circle (**CI**) cross (**CR**) line (**L**) block (**B**)

- <span>Type</span> 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

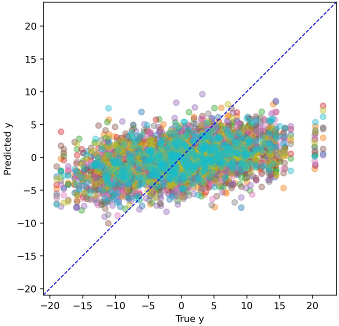
Numerical Illustration - Fitting, 60 × 60 Block Setting

(a)

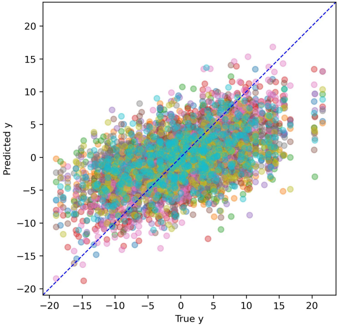


(a)True coefficient, $\mathcal{B}_0$. (b)-(c) Actual data (horizontal axis) against the predicted data (vertical axis) using $L = 10$ independent projection matrices of the same random projection type (colors). Training $n = 1000$, compression rate: $r = 0.36$, sparsity parameter $\psi = 3$.

(b) Tensor-wise



(c) Mode-wise (2)



Simulation studies: forecast fitting

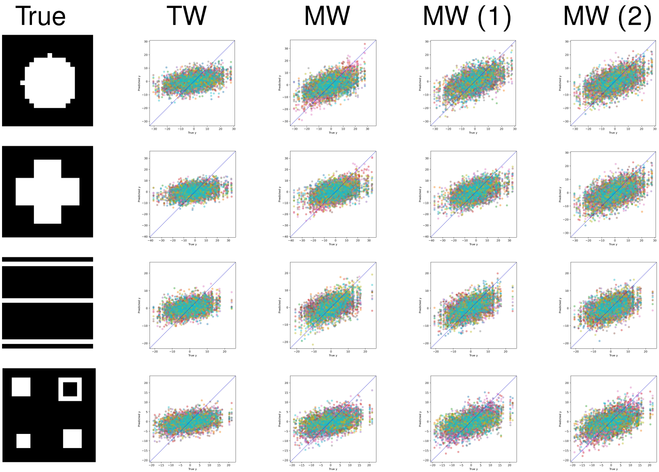


Figure: **Simulation results**: actual data against the predicted for different levels of sparsity (rows) and different types of random projections (columns), using 10 independent projection tensors (colours). For each plot: training sample size: $n = 1000$, compression rate: 0.36, $\psi = 3$.

Simulation studies: RMSE

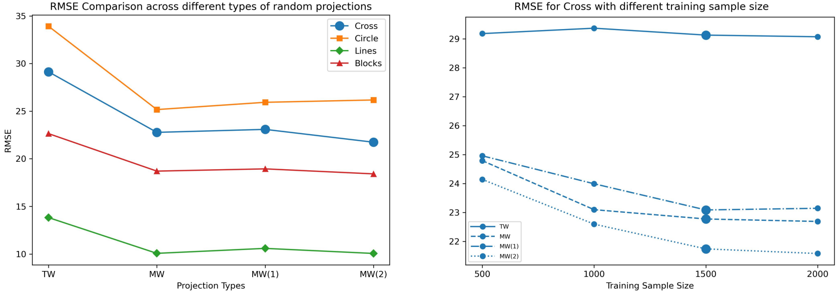


Figure: RMSE comparison across different types of random projection and different configurations in the baseline setting (top) and different sample sizes (bottom) in the 60×60 dimension case. Each estimate is obtained BMA over $L = 10$ independent projection matrices and 500 data points from the validation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

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

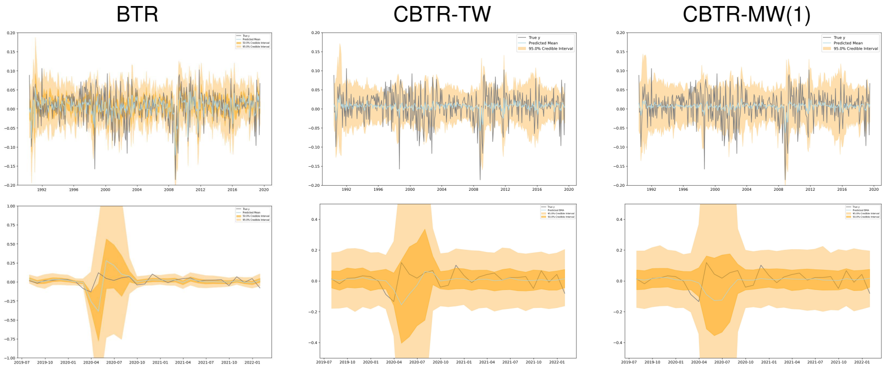
Specification

- Different from Xiao and Wang (2022), we consider Mixed Data Sampling (Rodriguez and Puggioni, 2010).
- y_t is the monthly 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: $t - 1/22, t - 2/22, \ldots, t - 22/22$.
- $\mathcal{X}_t \in \mathbb{R}^{7\times 22\times 4}$: variables × daily data × monthly lags.
- Training sample size $n = 350$.

$$y_t = \mu + \sum_{i_3=1}^{4} \left\langle B_{\tilde{I}(i_3)}, \begin{pmatrix} \mathrm{GV}_{t-\frac{1}{22}-i_3+1} & \mathrm{GV}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{GV}_{t-\frac{21}{22}-i_3+1} & \mathrm{GV}_{t-i_3} \\ \mathrm{BV}_{t-\frac{1}{22}-i_3+1} & \mathrm{BV}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{BV}_{t-\frac{21}{22}-i_3+1} & \mathrm{BV}_{t-i_3} \\ \mathrm{ER}_{t-\frac{1}{22}-i_3+1} & \mathrm{ER}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{ER}_{t-\frac{21}{22}-i_3+1} & \mathrm{ER}_{t-i_3} \\ \mathrm{IR}_{t-\frac{1}{22}-i_3+1} & \mathrm{IR}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{IR}_{t-\frac{21}{22}-i_3+1} & \mathrm{IR}_{t-i_3} \\ \mathrm{VI}_{t-\frac{1}{22}-i_3+1} & \mathrm{VI}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{VI}_{t-\frac{21}{22}-i_3+1} & \mathrm{VI}_{t-i_3} \\ \mathrm{TB}_{t-\frac{1}{22}-i_3+1} & \mathrm{TB}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{TB}_{t-\frac{21}{22}-i_3+1} & \mathrm{TB}_{t-i_3} \\ \mathrm{BD}_{t-\frac{1}{22}-i_3+1} & \mathrm{BD}_{t-\frac{2}{22}-i_3+1} & \cdots & \mathrm{BD}_{t-\frac{21}{22}-i_3+1} & \mathrm{BD}_{t-i_3} \end{pmatrix} \right\rangle + \sigma\varepsilon_t, \qquad (6)$$

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

Empirical application


BTR

CBTR-TW

CBTR-MW(1)

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	MW(1)	MW(1, 2)	MW(1, 3)	MW(2, 3)
In-sample	0.0338	0.0355	0.0346	0.0356	0.0333	**0.0323**	0.0329
Out-sample	0.1148	0.0676	0.0623	0.0723	**0.0383**	0.0600	0.0508

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.

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

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.

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.

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

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

Charikar, M., Chen, K., and Farach-Colton, M. (2004). Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15.

References II

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.

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*.

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

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.

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.

References III

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.

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.

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

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*.

References IV

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.

Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317.

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.

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

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.

Bayesian inference: convergence properties

Definition 1 (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 $\varepsilon > 0$.

Bayesian inference: convergence properties

Definition 1 (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 1 (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 $n \to \infty$.

Sketch of proof: Setup and Notation

- Tensor predictor: $\mathcal{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_D}$
- Compressed predictor: $\mathtt{GTRP}(\mathcal{X}_i)$
- Predictive density: $f(y \mid \langle \mathcal{B}, \mathtt{GTRP}(\mathcal{X}_i) \rangle)$
- 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

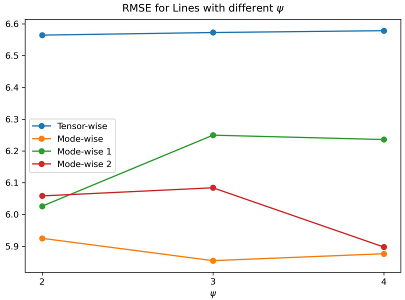
Goal: Show the prior puts enough mass near the true model f_0 by bounding

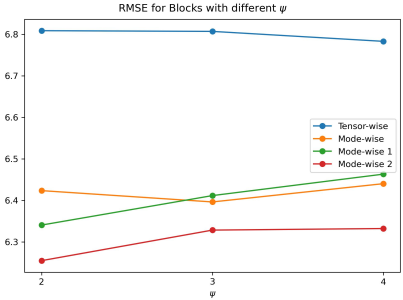
$$\pi\left(f : d(f, f_0) < \tfrac{1}{4}\varepsilon_n^2\right) \geq e^{-n\varepsilon_n^2/4}$$

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 $S = \left\{\mathcal{B} : |\langle \mathtt{GTRP}(\mathcal{X}_i), \mathcal{B}\rangle - \langle \mathcal{X}_i, \mathcal{B}_0\rangle| < \frac{\varepsilon_n^2}{4\eta}\right\}$
- $d_{t=1}(f, f_0) = \iint f_0\left(\frac{f_0}{f} - 1\right)\nu_y(dy)\nu_{\mathcal{X}}(d\mathcal{X}) =$ $E_{\mathcal{X}}\left[g(u^*)\left(\langle \mathtt{GTRP}(\mathcal{X}_i), \mathcal{B}\rangle - \langle \mathcal{X}_i, \mathcal{B}_0\rangle\right)\right]$.
- Choosing $|g(u^*)| < \eta$ implies that $d_t(f, f_0)$ is a subset of S, hence confirming condition 3.

Numerical Illustration - Sparsity

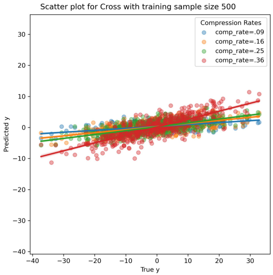


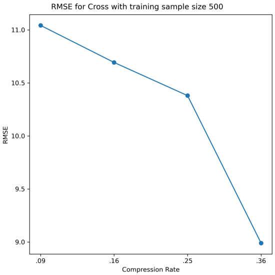


- Projections with different sparsity levels and random projection types: TW (blue), MW (orange), MW(1) (green), and MW(2) (red).
- $m = 500$ test samples, sparsity levels ($\psi \in \{2, 3, 4\}$) (horizontal axis).
- In most scenarios (**CI**, **CR**, and **L**), mode-wise random projection has the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

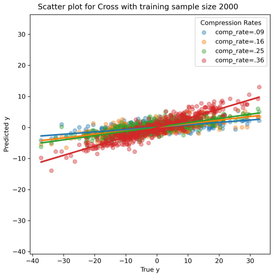
Numerical Illustration - Compression rate

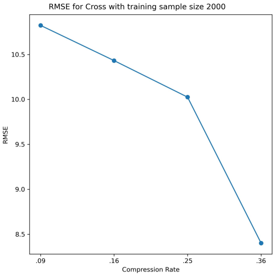
$n = 500$


Scatter plot for Cross with training sample size 500


RMSE for Cross with training sample size 500

$n = 2000$


Scatter plot for Cross with training sample size 2000


RMSE for Cross with training sample size 2000