

Compressed Bayesian Tensor Regression

Roberto Casarin[†], Radu Craiu^{‡*}, Qing Wang[§]

[†] Ca' Foscari University of Venice, Italy

[‡] University of Toronto, Canada

[§] Ca' Foscari University of Venice, Italy

Abstract

To address the common problem of high dimensionality in tensor regressions, we introduce a generalized tensor random projection method that embeds high-dimensional tensor-valued covariates into low-dimensional subspaces with minimal loss of information about the responses. The method is flexible, allowing for tensor-wise, mode-wise, or combined random projections as special cases. A Bayesian inference framework is provided featuring the use of a hierarchical prior distribution and a low-rank representation of the parameter. Strong theoretical support is provided for the concentration properties of the random projection and posterior consistency of the Bayesian inference. An efficient Gibbs sampler is developed to perform inference on the compressed data. To mitigate the sensitivity introduced by random projections, Bayesian model averaging is employed, with normalising constants estimated using reverse logistic regression. An extensive simulation study is conducted to examine the effects of different tuning parameters. Simulations indicate, and the real data application confirms, that compressed Bayesian tensor regression can achieve better out-of-sample prediction while significantly reducing computational cost compared to standard Bayesian tensor regression.

Keywords: Bayesian inference, posterior consistency, random projection, tensor regression

1 Introduction

Dimensionality reduction has been a key area of interest in learning from high-dimensional data. Traditional dimensionality reduction techniques, e.g., principal component analysis (PCA) and linear discriminant analysis, factor models, and sufficient dimensionality reduction, despite their effectiveness, suffer from severe computational restrictions which increase exponentially with the dimensions of the data (e.g., see Dasgupta, 2013, for a comparison between PCA and random projection).

In this paper, we consider random projection techniques, where randomly generated matrices are used to embed high-dimensional data points into a lower-dimensional space. Under fairly general assumptions, random projection preserves pairwise distances within a certain tolerance, as proved in the celebrated Johnson-Lindenstrauss (JL) lemma (Johnson and Lindenstrauss, 1984). Random projection has been successfully applied in statistics to reduce computational costs or to improve the efficiency of a standard method or model when applied to large datasets. For instance, Indyk and Motwani (1998); Ailon and Chazelle (2009); Datar et al. (2004) utilised it for the efficient approximation of the nearest neighbour search, Chakraborty (2023); Li et al. (2021); Cannings and Samworth (2017) applied it to high-dimensional classification, Dasgupta (1999) employed it to learn the mixture of Gaussian distributions in high dimensions, Li and Li (2023); Gondara and Wang (2020); Anagnostopoulos et al. (2018) used random projection to achieve data privacy, and Guhaniyogi and

*Corresponding author: radu.craiu@utoronto.ca (R. Craiu). Other contacts: r.casarin@unive.it (R. Casarin), qing.wang@unive.it (Q. Wang).

Dunson (2015); Farahmand et al. (2017); Koop et al. (2019) introduced random projection into inference for large dynamic regression models. In this paper, we focus on Bayesian tensor regression models, which have recently become popular in many fields for conducting inference and statistical learning based on multi-dimensional data (Guhaniyogi et al., 2017; Guhaniyogi, 2020; Billio et al., 2023, 2024; Luo and Griffin, 2025; Casarin et al., 2025). We consider scalar-on-tensor linear regressions, where dimensionality reduction is essential to reduce the number of parameters to estimate. In this sense, tensor decompositions have been used to extract factors from the covariate tensor or to parametrize the coefficient tensor in a hierarchical prior setting. However, when the number of covariates is so large that factors cannot be extracted optimally, then random projection offers a viable solution that is easy to implement and has strong theoretical guarantees in preserving the explanatory power of covariates.

Given the scarcity of literature on random projection within the Bayesian tensor regression framework, we contribute to this framework in several ways. Specifically, in this paper we i) extend the higher-order count sketch (HCS) method in Shi and Anandkumar (2019) and the projection technique in Li et al. (2021) to the case of tensor predictors; ii) provide concentration inequalities for the proposed projection; iii) integrate the projection into a tensor regression framework; iv) prove posterior consistency for the proposed compressed tensor regression; v) propose a Monte Carlo sampling procedure for posterior approximation under different prior specifications.

Different tensor random projection strategies have been studied in the literature. Rakhshan and Rabusseau (2020) proposed two types of tensorized random projections to map a mode- d tensor into a $q \times 1$ vector: $\mathbb{R}^{p_1 \times \dots \times p_d} \rightarrow \mathbb{R}^q$, using low-rank random projection tensors constructed by Tensor Train (Oseledets, 2011) or canonical polyadic (CP) representations such that each entry in $\mathbb{R}^q$ is computed from the inner product of a distinct random projection tensor and the tensor predictor. Shi and Anandkumar (2019) proposed an HCS that reduces the dimension of the original tensor while still preserving the higher-order data structure. In particular, given a 3-order tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, consider $\text{MTS}(\mathcal{X}): \mathbb{R}^{p_1 \times p_2 \times p_3} \rightarrow \mathbb{R}^{q_1 \times q_2 \times q_3}$, by taking the n -mode product along each mode of the tensor with a random hash matrix $H_m \in \mathbb{R}^{p_m \times q_m}$, $m = 1, 2, 3$, where q_1, q_2, q_3 are much smaller than p_1, p_2, p_3 . Their method is an extension of the count sketch Charikar et al. (2004). In a similar fashion, Li et al. (2021) proposed a random projection of a tensor by exploiting its CP representation, where the random projection is performed by randomly projecting each margin from the CP decomposition to a lower dimension. In this paper, we extend the projection proposed by Shi and Anandkumar (2019) and Li et al. (2021) to general order tensors, also allowing simultaneously for different projection strategies. Some modes of the tensor are projected separately into a lower space, whereas other modes are projected jointly, thus allowing for a reduction in the number of modes. We also derive JL-type concentration inequalities for the proposed tensor projection.

The JL lemma asserts that any set of n points in the d -dimensional Euclidean space can be embedded into the k -dimensional Euclidean space such that all pairwise distances are preserved within an arbitrarily small factor, $\epsilon > 0$, for $k = \mathcal{O}(\epsilon^{-2} \log n)$. The original JL lemma has been studied and proved in many ways to achieve faster embedding and tighter bounds (see Dasgupta and Gupta (2003) for a short and elegant proof of the original lemma). Central to the JL embedding is a $k \times d$ random projection matrix Φ . The original recipe requires Φ to meet three properties, namely, spherical symmetry, orthogonality, and normality (Ailon and Chazelle, 2009). These can be achieved by drawing each entry of Φ independently from a standard normal distribution, orthogonalizing each row using the Gram-Schmidt algorithm, and then normalising them to unit length. However, the resulting matrix is a dense matrix, which can slow down the evaluation of the random projection when the data dimension is large.

This motivates several variants of JL embeddings to simplify and sharpen the lemma. Indyk and Motwani (1998) showed that the JL guarantee can still be obtained without en-

forcing orthogonality and normality. Achlioptas (2003) not only dropped the spherical symmetry condition, but also proposed a sparse way to construct the random projection matrix. Each entry is independently drawn from a discrete distribution with atoms $-\sqrt{\psi}$, 0, and $\sqrt{\psi}$ with probability $1/2\psi$, $1 - 1/\psi$, and $1/2\psi$ where $\psi = 1$ or $\psi = 3$. To encourage sparsity in the random projection matrix and speed up computation, Li et al. (2006) used $\psi \gg 3$ (e.g., $\psi = \sqrt{D}$, where D is the number of features, or covariates). Matoušek (2008) considered a version of the JL lemma with independent sub-Gaussian projection entries. In this paper, we use tensor projections where the entries of the projection matrices and tensors are i.i.d. from the distribution used in Achlioptas (2003) and obtain new JL-type concentration inequalities by exploiting some properties of the Meijer G function in a significant departure from the existing literature (Mathai et al., 2010; Stojanac et al., 2018).

Random projections have also been used in Bayesian inference. For example, Chakraborty (2023) built an efficient Bayesian high-dimensional classifier using the same random projection as in Li et al. (2006), and Geppert et al. (2017) used random projection for Bayesian regression analysis. While their methods compress both the sample size and number of regressors, Guhaniyogi and Dunson (2015) proposed a compressed regression model where covariates are projected through $m \times p$ projection matrices with independent entries drawn from a discrete distribution with atoms $-\sqrt{\psi}$, 0 and $\sqrt{\psi}$ and probabilities $1/\psi^2$, $2(1 - 1/\psi)/\psi$ and $(1 - 1/\psi)^2$, respectively. To reduce the sensitivity on the choice of (m, ψ) values, Bayesian model averaging is used to average the results from s random projection matrices with different (m, ψ) values. Mukhopadhyay and Dunson (2020) generated the random projection matrix using a Targeted Random Project technique in which the probability of setting the j -th column to zero is proportional to the marginal dependence between predictor x_j and response variable y . In this paper, we extend the Bayesian compressed regression model to tensor regressions and provide some posterior consistency guarantees, building on the general consistency results that were derived in Jiang (2007).

The paper is organised as follows. Section 2 introduces the compressed tensor regression model as well as the probabilistic bounds for the tensor random projection. Section 3 presents the Gibbs sampler for sampling the tensor coefficients. Section 4 presents theoretical properties of posterior consistency for the coefficient posterior. The proofs for all the theoretical results are included in the Appendix. Section 5 presents the simulation results and a real-world dataset application. Section 6 concludes.

2 A Compressed Bayesian Tensor Model

2.1 Tensor random projection

A compressed Bayesian tensor regression (CBTR) model has the form

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

$j = 1, \dots, n$, where $\mu \in \mathbb{R}$ is the intercept, $\mathcal{B} \in \mathbb{R}^{q_1 \times \dots \times q_M}$ is the coefficient tensor, $\mathcal{X}_j \in \mathbb{R}^{p_1 \times \dots \times p_N}$ is the covariate tensor for the j th observation, and $\langle \cdot, \cdot \rangle$ is the scalar product for tensors (Kolda and Bader, 2009). $\text{GTRP}(\mathcal{X}_j)$ denotes the **Generalized Tensor Random Projection** (GTRP) operator applied to $\mathcal{X}_j$ defined as

$$\text{GTRP}(\mathcal{X}) := \mathcal{X} \times_1 H_1 \times_2 \dots \times_R H_R \times_{R+1:N} \mathcal{H}_{R+1:N}, \quad (2)$$

where $\mathcal{X} \in \mathbb{R}^{p_1 \times \dots \times p_N}$, where $\times_n$ and $\times_{n:m}$ denote the n -mode and the n -to- m mode products (Kolda and Bader, 2009), $H_m \in \mathbb{R}^{q_m \times p_m}$, $m = 1, \dots, R$ and $\mathcal{H} \in \mathbb{R}^{q_{R+1} \times \dots \times q_M \times p_{R+1} \times \dots \times p_N}$ are the random projection matrices and M -mode random projection tensor, respectively, with $R < M \leq N$. Without loss of generality, we assumed mode-wise projection for the first R modes, since the mode ordering can be chosen by the researcher. The GTRP proposed in Eq. (2) generalizes in two aspects the existing random projections for tensors. First, it extends

the projection for 3-mode tensors to tensors with a general number of modes N . Secondly, the projection reduces the dimensions of the covariate space, allowing for a smaller number of covariates within each mode, as well as a smaller number of modes. We define two distinct types of random projections used to construct our **GTRP**. The first type combines covariates of a given mode, while preserving the elements in the other modes. For that given mode, it is similar to classical techniques used in regression models, where new linear combinations of covariates are created to reduce collinearity.

Definition 1. A random projection **GTRP-MW** is called mode-wise when $\text{GTRP-MW}(\mathcal{X}) := \mathcal{X} \times_m H_m$ where $\mathcal{X} \in \mathbb{R}^{p_1 \times \dots \times p_N}$ and $H_m \in \mathbb{R}^{q_m \times p_m}$.

The second type uses the entries of a sub-tensor of $\mathcal{X}$ to obtain linear combinations conditionally independent given $\mathcal{X}$.

Definition 2. A random projection **GTRP-TW** is called tensor-wise when $\text{GTRP-TW}(\mathcal{X}) := \mathcal{X} \times_{n:m} \mathcal{H}$ where $\mathcal{X} \in \mathbb{R}^{p_1 \times \dots \times p_N}$ and $\mathcal{H} \in \mathbb{R}^{q_1 \times \dots \times q_M \times p_n \times \dots \times p_m}$, $M \leq N$ and $1 \leq n \leq m \leq N$.

It is apparent that $\text{GTRP-MW}(\mathcal{X})$ effectively changes the size of mode m from p_m to q_m , while still keeping the N -mode structure of $\mathcal{X}$, whereas $\text{GTRP-TW}(\mathcal{X})$ can be used to change either the number of modes or sizes of modes, or both. To gain an intuition of the **GTRP**, we consider some special cases that can serve as reference:

- (a) If $R = 0$, $M = 1$, **GTRP** corresponds to the random projection from N th-order tensor to q_1 dimensional vector: $\mathbb{R}^{p_1 \times \dots \times p_N} \rightarrow \mathbb{R}^{q_1}$. This setting doesn't exploit the original multiple-mode data structure and it is equivalent to the random projection in Achlioptas (2003) with $d = p_1 \times \dots \times p_N$ and $k = q_1$ applied to the vectorized tensor.
- (b) If $R = 0$, $M \geq 1$, only $\text{GTRP-TW}(\mathcal{X})_{i_1, \dots, i_M} = \langle \mathcal{X}, \mathcal{H}_{i_1, \dots, i_M, :} \rangle$ 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 modes will also be reduced from N to M .
- (c) If $R > 0$, $N = M = R + 1$, only $\text{GTRP-MW}(\mathcal{X})$ 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 **GTRP** involves both mode-wise random projection for the first R modes and tensor-wise random projection for the $(R + 1)$ th to N th modes. Similarly, mode reduction can be performed by choosing $M < N$.

To illustrate the effect of the mode preservation within our general **GTRP**, the following 2-mode covariate example, with one of the projection matrices being the identity, provides some insights.

Example 1. Considering a mode-wise random projection for a 3×2 matrix $\mathcal{X}$, $f(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 H_2$, where H_1 is a 3×3 identity matrix, H_2 is a 1×2 random row vector, this will map $\mathcal{X}$ into a 3×1 vector $\mathbf{x}$ with the entries,

$$\mathbf{x}_{i_1, i_2} = \sum_{j_1=1}^3 \sum_{j_2=1}^2 \mathcal{X}_{j_1, j_2} H_{1, i_1, j_1} H_{2, i_2, j_2} = \sum_{j_1=1}^3 \sum_{j_2=1}^2 \mathcal{X}_{j_1, j_2} \delta(j_1 = i_1) H_{2, i_2, j_2} = \sum_{j_2=1}^2 \mathcal{X}_{i_1, j_2} H_{2, i_2, j_2}.$$

Since the random projection matrix H_1 is the identity matrix, consistently with the definition of **GTRP-MW**, the random projection will only be performed in the second mode, thus returning a vector where the i_1 -th component is a linear combination of the elements of the i_1 -th row of $\mathcal{X}$.

As shown in the above illustrations, the value of R controls the extent of mode-wise random projection. The choice of using solely mode-wise random projection, tensor-wise random projection, or a combination of the two should be evaluated based on specific application requirements, as discussed in the numerical illustration section. Also, a trade-off between model performance and computational cost may be considered. In cases when dealing with very high-dimensional data with a large number of modes, a mode reduction can be performed by choosing $M < N$ to achieve computational feasibility. In contrast, when preserving the structural information is deemed necessary, the number of modes can remain unchanged by choosing $M = N$, while only reducing the dimensions along each mode.

Alternative random projections can be used. For instance, CP, TT and Kronecker Product (KP) decompositions can be applied with a given rank D to generate low-rank random projection tensors.

Example 2. *Considering random projections using the CP and TT methods in Rakhshan and Rabusseau (2020) to map a $p_1 \times p_2$ matrix $\mathcal{X}$ into a vector $\mathbf{x}$ as follows:*

$$\text{CPRP}(\mathcal{X})_i = \left\langle \sum_{d=1}^D A_{i,:,d}^1 \circ A_{i,:,d}^2, \mathcal{X} \right\rangle, \quad \text{TTRP}(\mathcal{X})_i = \langle \mathcal{G}_i^1 \times \mathcal{G}_i^2, \mathcal{X} \rangle, \quad (3)$$

where $A_i^n \in \mathbb{R}^{p_n \times D}$, $n=1,2$ and $\mathcal{G}_i^1 \in \mathbb{R}^{1 \times p_1 \times D}$ and $\mathcal{G}_i^2 \in \mathbb{R}^{D \times p_2 \times 1}$, $i = 1, \dots, q_1$.

Example 3. *Building on the Kronecker Product (KP) models introduced by Feng and Yang (2024) and on the relationship between KP and CP given in Batselier and Wong (2017), the CPRP and previous example projections can be extended to a Deep Kronecker random projection (DKRP). The DK definition and its relationship with the CP are*

$$\text{DKRP}(\mathcal{X})_i = \left\langle \sum_{d=1}^D \bigotimes_{\ell=1}^L \mathcal{B}_\ell^d, \mathcal{X} \right\rangle = \text{CPRP}(\mathcal{T}(\mathcal{X}))_i, \quad (4)$$

$i = 1, \dots, q_1$, where $\mathcal{T}$ is a one-to-one reshaping operator and CPRP used the tensor $\sum_{d=1}^D \bigcirc_{\ell=1}^L \text{vec}(\mathcal{B}_\ell^d)$. The operator $\mathcal{T}$ not only permutes the mode elements but also returns a tensor with a different number of modes.

Note that our mode-wise random projection can be thought of as constructing CP random projection tensors with rank 1 for each embedded entry. More importantly, CP, TT, and KP random projections map an order- N tensor to a vector that collapses all structural information; however, our methods still preserve the tensor structure, which can be valuable for practical applications.

A wide variety of distributions can be used for constructing the random projection matrices or tensors, provided that the entries are iid with mean zero and finite fourth moment (Mukhopadhyay and Dunson, 2020). A simple way to generate projections is to assume the elements of H_m and $\mathcal{H}_{R+1:N}$ are i.i.d. from a standard normal distribution. Dasgupta and Gupta (2003) gives concise proof of the JL lemma under the assumption of standard Gaussian entries. Nevertheless, the dense projection matrix used in classical random projection is not well-suited for high-dimensional problems. Thus, sparse and very sparse random projections have been proposed. In more applied literature, the Rademacher distribution is used in Rakhshan and Rabusseau (2021), to encourage sparsity in the constructed random projection matrices/tensors. In this paper, we follow Achlioptas (2003) and Li et al. (2006) and assume the entries are independent random variables from the following discrete distribution:

$$r = \sqrt{\psi} \begin{cases} +1 & \text{with prob. } \frac{1}{2\psi} \\ 0 & \text{with prob. } 1 - \frac{1}{\psi} \\ -1 & \text{with prob. } \frac{1}{2\psi} \end{cases} \quad (5)$$

2.2 Model properties

In our model the random projection $\text{GTRP}(\mathcal{X})$ projects the covariate tensor $\mathcal{X}_j \in \mathbb{R}^{p_1 \times \dots \times p_N}$ onto a lower-dimensional subspace that is: $\text{GTRP}(\mathcal{X}_j) \in \mathbb{R}^{q_1 \times \dots \times q_M}$, $j = 1, \dots, n$. The following results show that, when projecting, the distances between points in the original sample spaces are preserved by random projection under some suitable conditions. In the following, we define the constants $c(N, M) = p(N)/q(M)$, $p(N) = \prod_{m=1}^N p_m$, and $q(M) = \prod_{m=1}^M q_m$.

When $R = 0$ and $M = 1$, then $\text{GTRP}(\mathcal{X}_j)$ randomly projects all tensor entries into a vector space and the following JL concentration inequality holds uniformly in both the number of elements in each mode and in the number of modes.

Proposition 1 (A JL inequality for tensor-wise random projection). *Let $\mathbb{X}$ be an arbitrary set of n order N tensors in $\mathbb{R}^{p_1 \times \dots \times p_N}$. Define $\text{GTRP-TW}(\mathcal{X}) = \mathcal{X} \times_{1:N} \mathcal{H}_{1:N}$ with $\mathcal{H}_{1:N}$ an $N+1$ order random tensor in $\mathbb{R}^{p_1 \times \dots \times p_N \times q_1}$ with entries from the distribution in (5), and the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N, M)} \text{GTRP-TW}(\mathcal{X})$ from $\mathbb{R}^{p_1 \times \dots \times p_N}$ 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 $\mathcal{U}, \mathcal{V} \in \mathbb{X}$:*

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

The proof of Prop. 1 follows immediately from the proof of (Achlioptas, 2003, Thm.1.1), as the GTRP-TW is equivalent to the random projection in Achlioptas (2003). Additional details are provided in Appendix A.1.

Similarly, a concentration inequality can be proved when projecting mode-wise, that is, $R = M - 1, M = N$. The concentration bound is uniform in the number of elements in each mode but not in the number of modes.

Theorem 1 (JL inequality for mode-wise random projection). *Let $\mathbb{X}$ be an arbitrary set of n order N tensors in $\mathbb{R}^{p_1 \times \dots \times p_N}$. Let $\epsilon, \beta > 0$ and set*

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

Assume a sequence of positive integers q_j $j = 1, \dots, N$ satisfy $q(N) \geq q_0$ with probability at least $1 - n^{-\beta}$.

Define $\text{GTRP}(\mathcal{X}) = \mathcal{X} \times_1 H_1 \times_2 \dots \times_N H_N$, where the entries of $H_m \in \mathbb{R}^{p_m \times q_m}$ for $m = 1, \dots, N$ are independently distributed following the distribution given in (5), and the multilinear mapping $f(\mathcal{X}) = \sqrt{c(N, M)} \text{GTRP}(\mathcal{X})$ from $\mathbb{R}^{p_1 \times \dots \times p_N}$ to $\mathbb{R}^{q_1 \times \dots \times q_N}$. Then for all $\mathcal{U}, \mathcal{V} \in \mathbb{X}$, f satisfies

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

Theorem 1 extends classical JL inequalities from vectors to multiple-mode tensors that are projected along each mode. It also provides a theoretical foundation for using structured random projection for scalable Bayesian tensor regression. Note that setting $N = 1$ in the previous theorem yields the JL inequality from Proposition 1.

To get JL-embedding, we need that for each of the $\binom{n}{2}$ pairs of $\mathcal{U}, \mathcal{V} \in \mathbb{X}$, the squared norm of $(\mathcal{U} - \mathcal{V})$ is maintained within a factor of $1 \pm \epsilon$. If we can show that for some $\beta > 0$ and any fixed tensor $\mathcal{A} \in \mathbb{R}^{p_1 \times \dots \times p_N}$,

$$\Pr[(1 - \epsilon)\|\mathcal{A}\|^2 \leq \|f(\mathcal{A})\|^2 \leq (1 + \epsilon)\|\mathcal{A}\|^2] \geq 1 - \frac{2}{n^{2+\beta}}$$

then, by union bound, the probability of not getting a JL-embedding is bounded by $\binom{n}{2} \times \frac{2}{n^{2+\beta}} < \frac{1}{n^\beta}$.

A comparison of the bounds obtained from tensor-wise and mode-wise random projections is shown in Fig.1. The results illustrate the trade-off between maintaining the original data

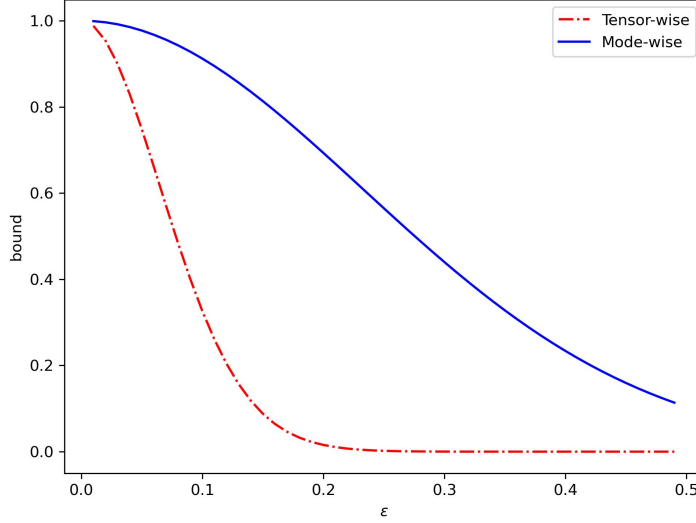


Figure 1: The plot shows the two bounds obtained by tensor-wise random projection according to Corollary 1 (red curve) and mode-wise random projection (blue curve) according to Theorem 1. We considered a 3-mode $\mathbb{R}^{20 \times 60 \times 50}$ tensor (i.e., $N = 3$, $p_1 = 20$, $p_2 = 60$ and $p_3 = 50$) projected into a $\mathbb{R}^{480}$ vector and a $\mathbb{R}^{4 \times 12 \times 10}$ tensor (i.e., $M = N = 3$, $q_1 = 4$, $q_2 = 12$ and $q_3 = 10$) with the mode-wise and tensor-wise projection, respectively, assuming $n = 10^4$ data points, and a concentration rate $\beta = 0.2$.

structure, such as some of the tensor modes, and the dimensions of the random subspace. In the case where all modes are preserved, the dimensionality reduction (blue line) is less effective than the case where the original structure vanishes completely (red line). Two main advantages of preserving the covariates' structure are the interpretability of projected covariates and the reduced computational cost.

The bounds presented above are optimal since they have been derived following a Chernoff-Cramér procedure. Alternative concentration bounds for our projections can be derived to provide some guarantees on the distance preservation. For example, based on a general hypercontractivity result of the Hanson-Wright type for polynomials of Gaussian and Rademacher variables (Hanson and Wright, 1971; Rakhshan and Rabusseau, 2020) and bounds on the moments up to the fourth-order, one can show the following bounds.

Theorem 2 (Alternative bounds using hyper-contractivity). *The JL-embedding can be achieved for GTRP with mode-wise random projection if $q(N) \geq q_0$, such that*

$$q_0 > C\epsilon^{-2}3^N(2+\beta)^{2N}\log^{2N}n, \quad (6)$$

with C an absolute constant.

Remark 1. *For the CP and TT projections, the following bounds on embedding dimensions have been obtained in Rakhshan and Rabusseau (2020)*

$$q_0 > C'\epsilon^{-2}3^{N-1}\left(1+\frac{2}{R}\right)^N\log^{2N}\left(\frac{n^{2+\beta}}{2}\right) \quad (7)$$

$$q_0 > C''\epsilon^{-2}\left(1+\frac{2}{R}\right)^N\log^{2N}\left(\frac{n^{2+\beta}}{2}\right), \quad (8)$$

where R denotes the rank of the random projection tensor, and C' and C'' are absolute constants.

The bounds given above are exponential and apply to general projection tensors even when the entries are not normally distributed. While they provide a Chernoff-like estimate,

the bounds are not optimal in the Chernoff-Cramér sense. We also note that they depend on absolute constants that are not easy to compute. The bounds in this section provide a theoretical basis for the methodological developments proposed in this paper. We emphasize that the bounds in Figure 1 are derived under general assumptions about the covariate tensor $\mathcal{X}$ and thus provide conservative bounds for those cases where $\mathcal{X}$ exhibits a more restrictive structure, such as high sparsity, or sparsity aligned with several coordinates. The difference in performance between tensor- and mode-wise projections, as suggested by Figure 1, has been confirmed by our numerical experiments in Section 5 and depends on the sparsity pattern in the covariate tensor.

2.3 Prior distributions

We consider two alternative specifications for the prior. In the first one, we assume independent Gaussian and inverse gamma prior distributions.

$$\mathcal{B} \sim \mathcal{TN}_{p_1, \dots, p_M}(\mathbf{0}, \Sigma_1, \dots, \Sigma_M), \quad \mu \sim \mathcal{N}(0, \sigma_\mu^2), \quad \sigma^2 \sim \mathcal{IG}(a, b). \quad (9)$$

In the second specification, we assume a hierarchical prior structure which builds on, as in Guhaniyogi et al. (2017), a Parallel Factor (PARAFAC) representation of $\mathcal{B}$ for further dimensionality reduction on tensor coefficients:

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

where $\circ$ denotes the *external product* of vectors, and $\gamma_m^{(d)}$ are the margins from PARAFAC decomposition of tensor coefficient $\mathcal{B}$. At first level, we assume that the margins from the PARAFAC decomposition are independent and follow multivariate normal distributions with zero mean vector and scales given by the product of the scalars τ , $\zeta^{(d)}$, and the diagonal matrix $W_m^{(d)} = \text{diag}(w_{m,1}^{(d)}, \dots, w_{m,j_m}^{(d)}, \dots, w_{m,q_m}^{(d)})$, i.e.

$$\gamma_m^{(d)} \sim \mathcal{N}_{q_m}(\mathbf{0}, \tau \zeta^{(d)} W_m^{(d)}), \quad m = 1, \dots, M, \quad d = 1, \dots, D. \quad (10)$$

This random scale specification allows for shrinkage at different levels.

To complete the hierarchical prior, at the second level, we modify the priors from Guhaniyogi et al. (2017) and assume the following prior distributions for the scales.

$$\tau \sim \mathcal{IG}(a_\tau, b_\tau), \quad w_{m,j_m}^{(d)} \sim \mathcal{Exp}((\lambda_m^{(d)})^2/2) \quad (11)$$

$$\lambda_m^{(d)} \sim \mathcal{Ga}(a_\lambda, b_\lambda), \quad (\zeta^{(1)}, \dots, \zeta^{(D)}) \sim \mathcal{Dir}(\alpha, \dots, \alpha), \quad (12)$$

$m = 1, \dots, M$, $d = 1, \dots, D$ where $\mathcal{IG}(a, b)$, $\mathcal{Ga}(a, b)$, $\mathcal{Exp}(\lambda)$ and $\mathcal{Dir}(\nu_1, \dots, \nu_D)$ denote the Inverse Gamma, Gamma, Exponential and Dirichlet distributions, respectively. The only difference compare to Guhaniyogi et al. (2017) is assuming the prior distribution of global shrinkage parameter τ is an Inverse Gamma instead of Gamma, largely due to the fact that τ appears as a variance parameter in the Gaussian prior of $\gamma_m^{(d)}$, it is natural to assume $\tau \stackrel{\text{i.i.d.}}{\sim} \mathcal{IG}(a, b)$ to get a more tractable full conditional distribution in the posterior approximation procedure.

3 Posterior approximation

3.1 Gibbs sampling

The joint posterior distribution $f(\gamma_m^{(d)}, \zeta^{(d)}, \tau, \lambda_m^{(d)}, w_m^{(d)}, \sigma^2, \mu \mid \mathbf{y}, \text{GTRP}(\mathcal{X}))$ is not tractable, so it must be approximated using the Monte Carlo method. We achieve this using a custom-built Gibbs sampler. Below, we describe the conditional sampling steps required by the algorithm's design. The derivation of the full conditionals can be found in Appendix B.1. The sampler cycles between the following steps:

1. Draw $\gamma_m^{(d)}$ from a multivariate normal distribution $f(\gamma_m^{(d)} | \mathbf{y}, \text{GTRP}(\mathcal{X}), \gamma_{-m}, \tau, \zeta, \mathbf{w}, \mu, \sigma^2)$ for $d \in \{1, \dots, D\}$ and $m \in \{1, \dots, M\}$.

Let us denote the Generalized Inverse Gaussian distributions with $\mathcal{GIG}$. The Gibbs updates for the remaining parameters and hyper-parameters are:

2. Draw $\zeta^{(d)}$ from the $\mathcal{GIG}$ distribution $f(\zeta^{(d)} | \gamma^{(d)}, \tau, \mathbf{w}^{(d)})$.
3. Draw τ from the $\mathcal{IG}$ distribution $f(\tau | \gamma, \zeta, \mathbf{w})$.
4. Draw $\lambda_m^{(d)}$ from a Gamma distribution $f(\lambda_m^{(d)} | \gamma_m^{(d)}, \tau, \zeta^{(d)})$.
5. Draw $w_{m,j_m}^{(d)}$ from the $\mathcal{IG}$ distribution $f(w_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)})$.
6. Draw σ^2 from the $\mathcal{IG}$ distribution $f(\sigma^2 | \mathbf{y}, \text{GTRP}(\mathcal{X}), \mu, \gamma)$.
7. Draw μ from the Gaussian distribution $f(\mu | \mathbf{y}, \text{GTRP}(\mathcal{X}), \gamma, \sigma^2)$.

The full conditional distributions of the Gibbs sampler for the hierarchical Normal-Inverse Gamma prior are given in Appendix B.2. Both variants of the Gibbs algorithm involve conditional densities that are available in closed form and can be sampled exactly.

3.2 Model averaging

Reliance on a single random projection is a risky approach, since one may not be sure of the optimal type of projection or how far the projection matrix is from an optimal one. Moreover, it is straightforward to parallelise the computation and substantially reduce the time to obtain estimates or predictions from several projections. In this paper, we focus on prediction and propose to use Bayesian model averaging to combine the predictions produced by different compressed tensor regressions.

Specifically, we generate L different random projections for each compressed tensor regression using entries randomly drawn from the distribution proposed in (5). Let $\mathcal{M}_\ell, \ell = 1, \dots, L$, represent the model in (1) with $\text{GTRP}^{(\ell)}(\cdot)$ denoting the distinct random projection for $\mathcal{M}_\ell$. We further denote f_ℓ the predictive density for $\mathcal{M}_\ell$ and $\boldsymbol{\theta}^{(\ell)} = (\mu^{(\ell)}, \mathcal{B}^{(\ell)}, \sigma^{2(\ell)})$ its parameters, $\mathcal{D} = \{(y_j, \text{GTRP}(\mathcal{X}_j)), j = 1, \dots, n\}$ the observed data, and we are interested in the predictive density of $y_{n+j'}$ given $\mathcal{X}_{n+j'}$

$$f(y_{n+j'} | \text{GTRP}^{(\ell)}(\mathcal{X}_{n+j'}), \mathcal{D}) = \sum_{\ell=1}^L p_\ell(\mathcal{M}_\ell | \mathcal{D}) f_\ell(y_{n+j'} | \text{GTRP}^{(\ell)}(\mathcal{X}_{n+j'}), \mathcal{D}, \mathcal{M}_\ell) \quad (13)$$

$$f_\ell(y_{n+j'} | \text{GTRP}^{(\ell)}(\mathcal{X}_{n+j'}), \mathcal{D}, \mathcal{M}_\ell) = \int f_\ell(y_{n+j'} | \text{GTRP}^{(\ell)}(\mathcal{X}_{n+j'}), \boldsymbol{\theta}^{(\ell)}, \mathcal{M}_\ell) p_\ell(\boldsymbol{\theta}^{(\ell)} | \mathcal{M}_\ell, \mathcal{D}) d\boldsymbol{\theta}^{(\ell)} \quad (14)$$

for $j' = 1, \dots, m$ where m is the size of the validation set. Since the normalizing constant $c_\ell = p_\ell(\mathcal{M}_\ell | \mathcal{D})$ of $p_\ell(\boldsymbol{\theta}^{(\ell)} | \mathcal{M}_\ell, \mathcal{D})$ is not available in closed form, we approximate it using reverse logistic regression, as recommended by Geyer (1994).

To approximate the predictive density in (13), we first evaluate empirically the predictive distribution of $y_{n+j',s}^{(\ell)}$ produced by the ℓ th random projection. In particular, at the s th MCMC step a random draw $y_{n+j',s}^{(\ell)}$ is generated from the posterior predictive distribution

$$y_{n+j',s}^{(\ell)} | \text{GTRP}^{(\ell)}(\mathcal{X}_{n+j'}), \boldsymbol{\theta}_s^{(\ell)} \sim \mathcal{N} \left(\mu_s^{(\ell)} + \left\langle \text{GTRP}^{(\ell)}(\mathcal{X}_{n+j'}), \mathcal{B}_s^{(\ell)} \right\rangle, \sigma_s^{2(\ell)} \right), \quad (15)$$

where $\boldsymbol{\theta}_s^{(\ell)}, s = 1, \dots, S$ denote the MCMC draws from the posterior distribution.

We pool $y_{n+j',s}^{(\ell)}$ across ℓ and s to obtain an empirical distribution which approximates the distribution of $y_{n+j'}$. If $\tilde{y}_{n+j'}^{(\ell)}$ denotes the approximated prediction given $\mathcal{D}$ and the ℓ th projection $\mathbf{GTRP}^{(\ell)}(\mathcal{X}_{n+j'})$, we approximate the posterior predictive mean with

$$\tilde{y}_{n+j'} = \sum_{\ell=1}^L w_{\ell} \tilde{y}_{n+j'}^{(\ell)}, \quad \tilde{y}_{n+j'}^{(\ell)} = \frac{1}{S} \sum_{s=1}^S y_{n+j',s}^{(\ell)},$$

where $w_{\ell} = c_{\ell} / \sum_{k=1}^L c_k$, for all $\ell = 1, \dots, L$.

To evaluate the quantiles of the predictive distribution $f(y_{n+j'} \mid \mathbf{GTRP}(\mathcal{X}_{n+j'}, \mathcal{D}))$ define $z_{n+j',s} = \sum_{\ell=1}^L u_{n+j',s}^{(\ell)} y_{n+j',s}^{(\ell)}$, where $(u_{n+j',s}^{(1)}, \dots, u_{n+j',s}^{(L)}) \sim \text{Multinomial}(1, (w_1, \dots, w_L))$. Because

$$P(z_{n+j',s} \leq t) = \sum_{\ell=1}^L P(z_{n+j',s} \leq t \mid u_{n+j',s}^{(\ell)} = 1) P(u_{n+j',s}^{(\ell)} = 1) = \sum_{\ell=1}^L P(y_{n+j',s}^{(\ell)} \leq t) w_{\ell}$$

we have $f(t \mid \mathbf{GTRP}(\mathcal{X}_{n+j'}, \mathcal{D})) = \sum_{\ell=1}^L w_{\ell} f^{(\ell)}(t \mid \mathbf{GTRP}(\mathcal{X}_{n+j'}, \mathcal{D}), \mathcal{M}_{\ell})$. So the quantiles for the density f in (13) can be evaluated from the sample quantiles of the L predictive distributions defined in (14).

4 Posterior Consistency

Projection of the tensor predictor is justifiable from a computational point of view, but the statistical validity of the resulting inference must be defensible theoretically. To this end, we present in this section theoretical results that demonstrate that the predictions generated, respectively, with the original and compressed tensor predictors and variables can be made arbitrarily close for particular choices of the projection matrix.

4.1 Notation and background

To show the posterior consistency of the model predictions, we consider, without loss of generality, the modewise random projection of the 3-mode tensors to the 3-mode tensors with a reduced number of elements along the modes. Let $\mathcal{X}_j \in \mathbb{R}^{p_{1,n} \times p_{2,n} \times p_{3,n}}$ denote the 3-mode tensor predictor for observation $j = 1, \dots, n$. We assume that there is a true tensor coefficient $\mathcal{B}_0 \in \mathbb{R}^{p_{1,n} \times p_{2,n} \times p_{3,n}}$. Denote by $\mathbf{GTRP-M}(\mathcal{X}_j)$, $\mathcal{B} \in \mathbb{R}^{q_{1,n} \times q_{2,n} \times q_{3,n}}$ the compressed tensor predictor and coefficient, respectively. Let $p_n = p_{1,n} \times p_{2,n} \times p_{3,n}$ and $q_n = q_{1,n} \times q_{2,n} \times q_{3,n}$ denote the number of predictors for a given sample size n before and after compression, respectively.

Let f_0 be the true posterior predictive density given the predictors $\mathcal{X}$, and f be the predictive density given the coefficients $\mathcal{B}$ drawn from its posterior distribution and the predictors $\mathcal{X}$. Let $\nu_{\mathcal{X}}(d\mathcal{X})$ be the probability measure for $\mathcal{X}$, and $\nu_y(dy)$ be the dominating measure for conditional densities f and f_0 . We assume that the true relationship between the response y and the predictors $\mathcal{X}$ follows a parametric generalized linear model (GLM) of the form $f(y \mid \mathcal{X}, \mathcal{B}_0) = \exp\{a(h)y + b(h) + c(y)\}$, where $h = \langle \mathcal{X}, \mathcal{B}_0 \rangle$. In the case of a normal linear regression, with mean h and variance σ^2 , the density is obtained by choosing $a(h) = h/\sigma^2$, $b(h) = -h^2/(2\sigma^2) - 1/2 \ln(2\pi\sigma^2)$ and $c(y) = -y^2/(2\sigma^2) - 1$.

The following measures of closeness are used to show posterior consistency. The Hellinger distance between f and f_0 and the Kullback-Leibler divergence of f from f_0 are

$$d(f, f_0) = \sqrt{\iint (\sqrt{f} - \sqrt{f_0})^2 \nu_{\mathcal{X}}(d\mathcal{X}) \nu_y(dy)}$$

$$d_{KL}(f, f_0) = \iint f_0 \ln \left(\frac{f_0}{f} \right) \nu_{\mathcal{X}}(d\mathcal{X}) \nu_y(dy),$$

respectively. In addition, we define

$$d_t(f, f_0) = t^{-1} \left(\iint f_0 \left(\frac{f_0}{f} \right)^t \nu_{\mathcal{X}}(d\mathcal{X}) \nu_y(dy) - 1 \right), \quad \forall t > 0.$$

4.2 Posterior results

In this section, we present two important theoretical results on posterior consistency of CBTR using two different priors for the tensor coefficients $\mathcal{B}$: the Gaussian prior and the PARAFAC prior. The following theorems on consistency are proved by verifying that the sufficient conditions a, b and c in Theorem 4 of Jiang (2007) are satisfied. The theoretical results derived in this section rely on the following assumptions:

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

Assumption **A.1** imposes restrictions on the growth rate of the number of regressors, q_n , so that q_n grows sublinearly with the total number of observations. Intuitively, this assumption prevents the projected model from being “too” complex.

$$\mathbf{A.2} \quad \bar{\lambda}_n \leq B q_n^v, \quad \underline{\lambda}_n \geq B_1 (\log(q_n))^{-1} \text{ for some positive constants } B, B_1, v.$$

Assumption **A.2** imposes some constraints on the prior covariance matrix of $\mathcal{B}$ by bounding the eigenvalues of the covariance matrix to ensure that the prior is well-defined and does not allow it to be too diffuse or too concentrated. However, conditions in **A.2** are mild and can be easily met.

$$\mathbf{A.3} \quad \frac{\log(\|\text{GTRP}(\mathcal{X})\|)}{n\varepsilon_n^2} \rightarrow 0, \quad \|\text{GTRP}(\mathcal{X})\|^2 > 8 \frac{(K^2+1) \log(q_n)}{B_1 n\varepsilon_n^2}, \quad \forall \mathcal{X} = \mathcal{X}_1, \dots, \mathcal{X}_n.$$

Assumption **A.3** ensures that the tensor random projection operation $\text{GTRP}(\cdot)$ does not excessively distort the norm of the tensor covariates $\mathcal{X}$, thus preserving the power of the covariates to explain the responses. This assumption is typically satisfied with high probability for carefully designed random projections as described in Proposition 1 and Theorem 1.

$$\mathbf{A.4} \quad D(\log(\|\text{GTRP}(\mathcal{X}_i)\|) + \log D) \sum_{m=1}^M q_{m,n} < M n \varepsilon_n^2 C \text{ for some positive constant } C.$$

$$\mathbf{A.5} \quad \varepsilon_n^2 = n^\delta \text{ with } b-1 < \delta < 0 \text{ where } \sum_{m=1}^M q_{m,n} = \mathcal{O}(n^b).$$

Assumption **A.4** and **A.5** target PARAFAC priors on compressed tensor coefficients $\mathcal{B}$. Assumption **A.4** controls the complexity of the model by bounding the projection norm $\|\text{GTRP}(\mathcal{X}_i)\|$, the PARAFAC component D , and the number of coefficients $D \sum_{m=1}^M q_{m,n}$. The condition in **A.4** compresses both the entropy and the prior mass by reducing the number of parameters and limiting the parameter space. Assumption **A.5** mainly specifies how fast the posterior contracts, at a rate slower than n^{-1} , but still converging. It also controls the growth of the projected dimension: the total number of compressed parameters $q_{m,n}$ must grow sublinearly with n . Altogether, assumption **A.5** ensures that the predictive distribution does not overfit as n grows.

Theorem 3. Let $\mathcal{B} \sim \mathcal{TN}(\mathbf{0}, \Sigma_1, \dots, \Sigma_N)$ a priori and $\tilde{\lambda}_n$ and $\underline{\lambda}_n$ be the largest and smallest eigenvalues of $\Sigma_1, \dots, \Sigma_N$. In addition, assume that all the covariates are bounded, which means $|x_{jkl}| < 1$ and $\lim_{n \rightarrow \infty} \sum_{j=1}^{p_{1,n}} \sum_{k=1}^{p_{2,n}} \sum_{l=1}^{p_{3,n}} |b_{jkl,0}| < K$. Define $D(R) = 1 + R \sup_{|h| \leq R} |a'(h)| \sup_{|h| \leq R} \left| \frac{b'(h)}{a'(h)} \right|$, $\theta_n = \sqrt{q_n p_n}$. For a sequence ε_n satisfying $0 < \varepsilon_n^2 < 1$ and $n\varepsilon_n^2 \rightarrow \infty$, assume that the assumptions **A.1**, **A.2** and **A.3** hold, then

$$E_{f_0} \pi [d(f, f_0) > 4\varepsilon_n \mid (y_j, \mathcal{X}_j)_{j=1}^n] \leq 4e^{-n\varepsilon_n^2/2}, \quad (16)$$

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

Theorem 4. 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, that is, $|x_{jkl}| < 1$ and $\lim_{n \rightarrow \infty} \sum_{j=1}^{p_{1,n}} \sum_{k=1}^{p_{2,n}} \sum_{l=1}^{p_{3,n}} |b_{jkl,0}| < K$. For a sequence ε_n satisfying $0 < \varepsilon_n^2 < 1$ and $n\varepsilon_n^2 \rightarrow \infty$, assume that the assumptions **A.1**, **A.4** and **A.5** hold then

$$E_{f_0} \pi [d(f, f_0) > 4\varepsilon_n \mid (y_i, \mathcal{X}_j)_{j=1}^n] \leq 4e^{-n\varepsilon_n^2/2}, \quad (17)$$

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

5 Numerical Illustrations

5.1 Simulations

We performed simulations under different settings for the type of random projection (tensor-wise and mode-wise), covariate tensor dimensions (20×20 and 60×60 mode-2 tensors), and the number of observations (from 500 to 2000 at an interval of 500). In addition, we investigated the sensitivity to compression rate, defined as $r = 1/C(N, M)$, where we recall $C(N, M) = p(N)/q(M)$ with $p(N) = \prod_{m=1}^N p_m$, and $q(M) = \prod_{m=1}^M q_m$, and different values of the sparsity coefficient ψ used in generating projection matrices (tensors) and the PARAFAC decomposition rank.

The configurations of the tensor coefficient are presented in panel (a) of Figure 3 and are labeled circle (**CI**), cross (**CR**), line (**L**), and block (**B**). The **CI** and **CR** configurations are symmetric along all modes and are sparse with different sparsity levels. The **L** and **B** configurations are asymmetric along at least one mode and represent scenarios where projections that preserve the mode can improve the results. The tensor covariates are drawn independently from the standard normal distribution. The efficiency of Gibbs sampling has been proved computationally on a tensor regression model without projection (Casarin et al., 2025). See Appendix C for an illustrative example of MCMC output.

For each simulation setting, we performed $L = 10$ independent random projections of the same type and combined the results using Bayesian model averaging. This required 2560 simulations for a given ψ . We evaluated the performance of different models using posterior predictive checks. Several quantities are used to evaluate the model fitting. The distance of the actual data from their mean is defined as follows:

$$d_j = (y_j - \bar{y})^2, j = n+1, \dots, n+m, \quad \bar{y} = \frac{1}{m} \sum_{j=n+1}^{n+m} y_j. \quad (18)$$

The root mean square error across the L independent projections of the same type is defined as

$$RMSE_{j,n} = \sqrt{\frac{1}{L} \sum_{\ell=1}^L (y_j - \tilde{y}_{j,n}^{\ell})^2}, j = n+1, \dots, n+m, \quad (19)$$

where $\tilde{y}_{j,n}$ is the point prediction obtained for the j th out-of-sample item, based on a training sample of size n .

5.1.1 Type of projection

The top plots in Fig. 2 show the RMSE (vertical axis) for the different baseline settings where 20×20 (panel a) and a 60×60 (panel b) true tensor coefficients are used in generating $n = 1,500$ i.i.d. samples from the tensor-regression model. In each plot, the RMSEs are reported for each projection method (horizontal axis) and configuration setting (different lines and symbols). The tensor-wise projection (first symbol in the four lines) underperformed the mode-wise projections in our four simulation settings.

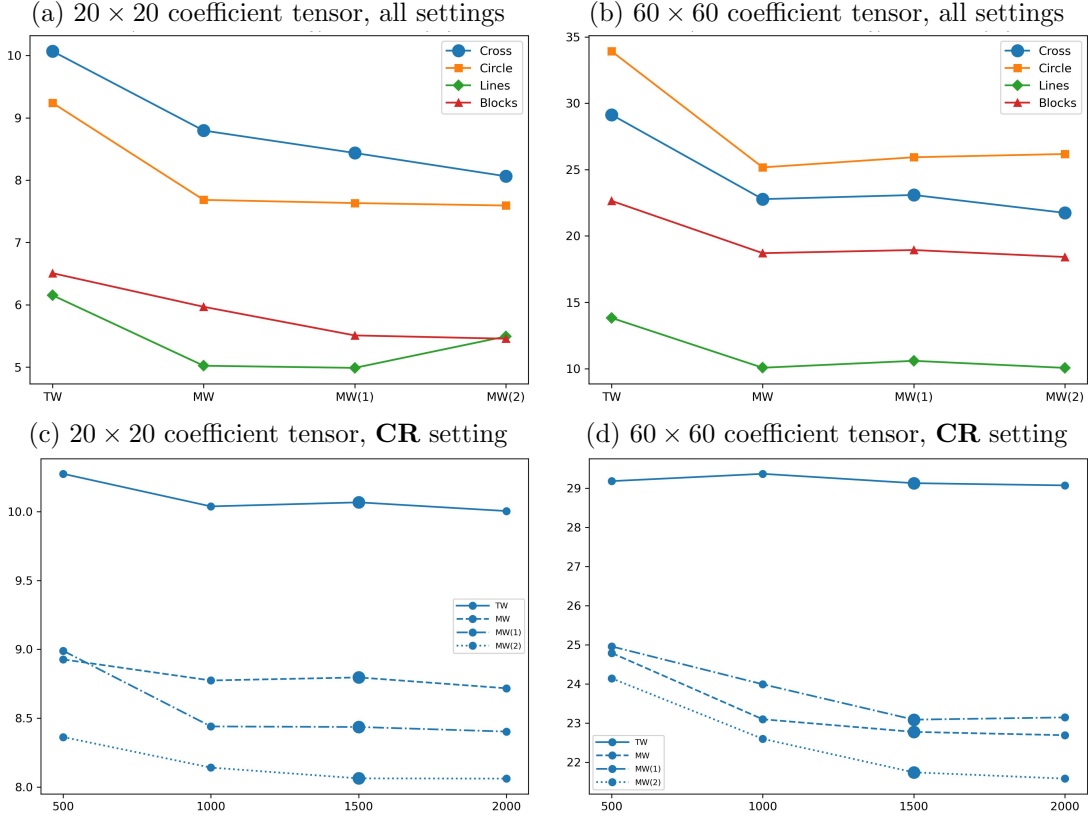


Figure 2: RMSE comparison across types of random projections (TW: tensor-wise, MW: mode-wise, MW(1): mode-wise preserving the first mode, and MW(2): mode-wise preserving the second mode), settings (blue: **CR**, orange: **CI**, green: **L** and red: **B**), and dimensions ((a): 20×20 and (b): 60×60). The top panels show the RMSEs (vertical axis) obtained for a training sample of size 1500 for different projection types (horizontal axis) in different settings (colors and symbols). The bottom panels show the RMSEs (vertical axis) for different training sample sizes (horizontal axis) and different projection types (line types). Each estimate is obtained via BMA over $L = 10$ independent projection matrices of the same type and 500 data points from the validation set. The larger dots in plots (c) and (d) indicate the RMSEs reported in the blue line of plots (a) and (b), respectively.

The bottom plots in Figure 2 show the RMSE (vertical axis) for different training sample sizes (horizontal axis) for the simulation setting **CR** with different types of random projections (different lines). As a reference, the larger dots in each line indicate the RMSEs reported in the blue line of panel (a). There is a clear downward-sloping trend as the training sample size increases across all random projection types, with the mode-wise projections outperforming the tensor-wise. Among the mode-wise projections, the one preserving the second mode performs best (dotted line).

We investigate the features of the different projection methods by comparing the actual values y_{n+j} in the test set with their predicted values $\tilde{y}_{n+j}^{(\ell)}$ (scatter plots in Fig. 3). Column 1 of panel (b) has been obtained using tensor-wise random projection (GTRP-TW) and Bayesian tensor regression on a training sample of $n = 1,000$ observations and a test sample of $m = 500$ observations. Compression rate $r = 0.36$ and sparsity coefficient $\psi = 3$ are used to generate the random projection tensors. Each plot reports the true (horizontal axis) and the predicted response variable (vertical axis). The estimation and prediction exercise has been performed using $L = 10$ independent projections of the same type (different colored dots) and different data-generating settings (different rows).

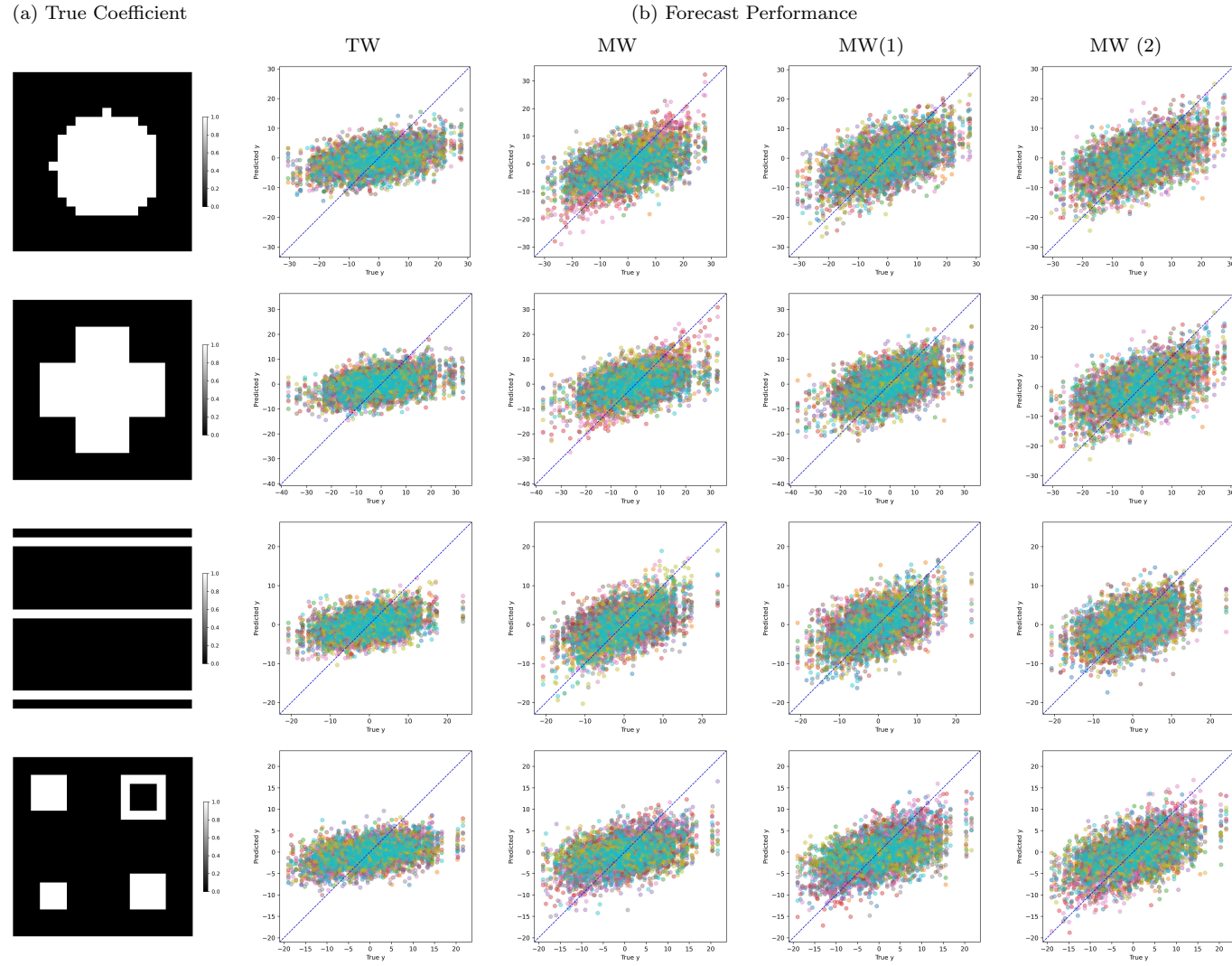


Figure 3: True coefficient (panel a) and forecast (panel b). In each scatter plot: actual data (horizontal axis) against the predicted data (vertical axis) for different sparsity levels and structures (rows) and different types of random projections (columns), using $L = 10$ independent projection matrices of the same random projection type (colors). In the experiments: training sample size $n = 1000$, compression rate: $r = 0.36$, sparsity parameter $\psi = 3$.

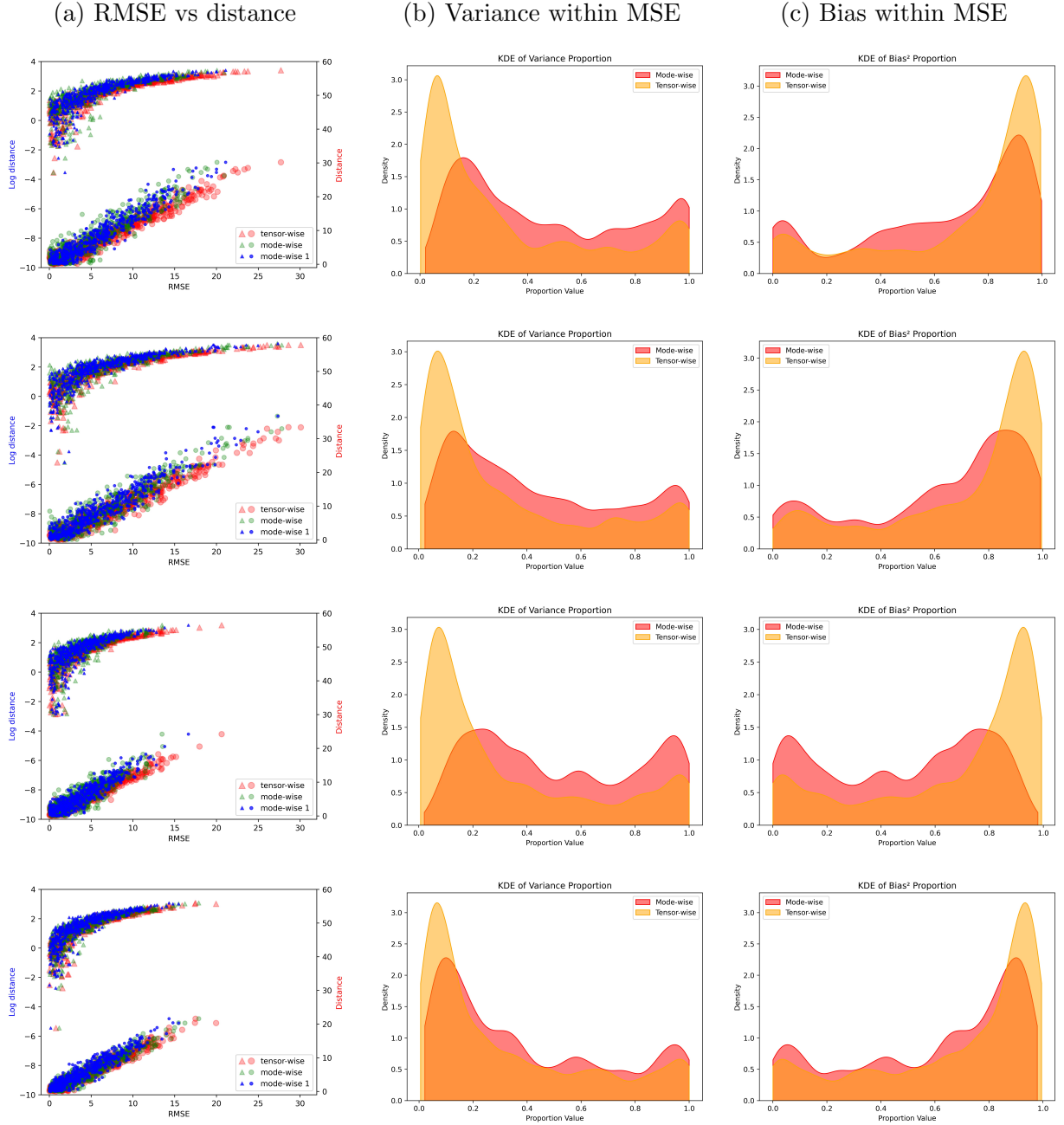


Figure 4: Prediction errors. Panel (a) shows RMSE vs actual distance d_j as defined in (18) (circle plotting symbols, right axis) and log-distance (triangle plotting symbols, left axis) between $m = 500$ data points and their mean obtained from different types of random projections: TW (pink), MW (green), and MW(1) (blue). Panels (b) and (c) show the decomposition of MSE obtained from the $m = 500$ test samples for two different types of random projections: TW (yellow) and MW (pink). Panel (b) shows the variance contribution to the MSE, and panel (c) shows the bias contribution to the MSE.

The same prediction evaluation has also been carried out for random projection types: Mode-wise (GTRP-MW), Mode-wise preserving mode 1 (GTRP-MW(1)), and Mode-wise preserving mode 2 (GTRP-MW(2)) (columns from 2 to 4, respectively). The plots in panel (b) show that GTRP-TW has difficulties in fitting the actual data (comparing the distance of the clouds from the 45° reference line). In contrast, GTRP-MW, GTRP-MW(1), GTRP-MW(2) perform better for values of the actual data both close and far from the mean.

To further investigate the relationship between the incurred errors and the relative distance of an observation from its distribution's mean, we produce graphical representations of

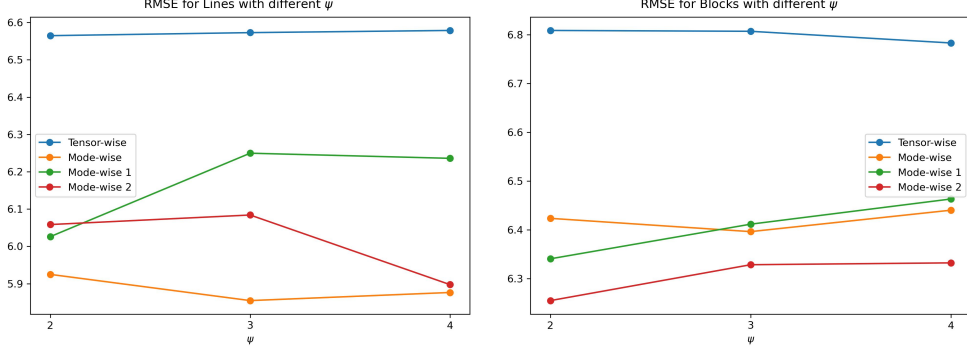


Figure 5: Effects of using random projection matrices of different sparsity levels on prediction errors (RMSE) for **L** (left) and **B** simulation settings (right). In the two plots: the RMSE (vertical axis) obtained from $m = 500$ test samples versus the sparsity levels ($\psi \in \{2, 3, 4\}$) (horizontal axis) for random projection types: TW (blue), MW (orange), MW(1) (green), and MW(2) (red).

the relationship between the distance d_j defined in (18) and the forecasting error $\text{RMSE}_{j,n}$ $j = n + 1, \dots, n + m$.

In the leftmost column of Figure 4 we show scatter plots of distance (marked with circles), and scatter plots of distances on the log-scale (marked with triangles) versus RMSE. We use two different scales for distances because we are interested in regions where distances are small (and the log scale explodes to $-\infty$) and regions in the right tail where the log scale is more interpretable. In every plot, the top cloud (triangle symbols) shows the tail behavior, while the bottom one (circle symbols) shows the relationship in the center of the distribution.

Blue symbols are generally at the left of the other color symbols, suggesting mode-wise random projection with mode-preserving yields smaller RMSE given the same distances.

The right column presents the empirical distribution of the variance and bias proportion of the m points of the test sample. The forecasts for the $m = 500$ points of the test sample are obtained with a training sample size $n = 1000$. The decomposition of MSE shows that tensor-wise random projection yields smaller variances but higher bias across all four different simulation settings than mode-wise random projection.

5.1.2 Sparsity and compression rates

Parameter ψ controls the sparsity level in the random projection tensor. When $\psi = 1$, the entries of the random projection tensor are essentially drawn from $\{-1, 1\}$ with equal probabilities (a Rademacher distribution), which is considered a non-sparse projection tensor. As the value of ψ increases, the entries of the random projection tensor will be drawn from $\{-1, 0, 1\}$ with increasing probability that 0 is drawn, and the projection tensor becomes sparser as ψ increases. For example, the probabilities of 0 being drawn are $1/2, 2/3, 3/4$ corresponding to ψ taking values 2, 3 and 4.

Fig. 5 reports the RMSE for simulation configurations of **L** and **B** using $\psi \in \{2, 3, 4\}$ representing dense to sparse random projection tensors (plots for other configurations can be found in the Supplementary Materials). The model averaging is performed across different random projections and different training sample sizes and the BMA's performance is evaluated using the RMSE. Fig. 5 suggests that tensor-wise random projection is not as sensitive as mode-wise random projection for varying sparsity of the random projection matrices. Mode-wise random projections and mode-wise random projections with mode preserving still outperform tensor-wise random projections. In most scenarios (**CI**, **CR**, and **L**), mode-wise random projection has the lowest RMSE compared to the other random projection methods. A V-shape curve is observed for mode-wise random projection, suggesting that a moderate sparsity in the random projection process is preferred and helps preserve more information.

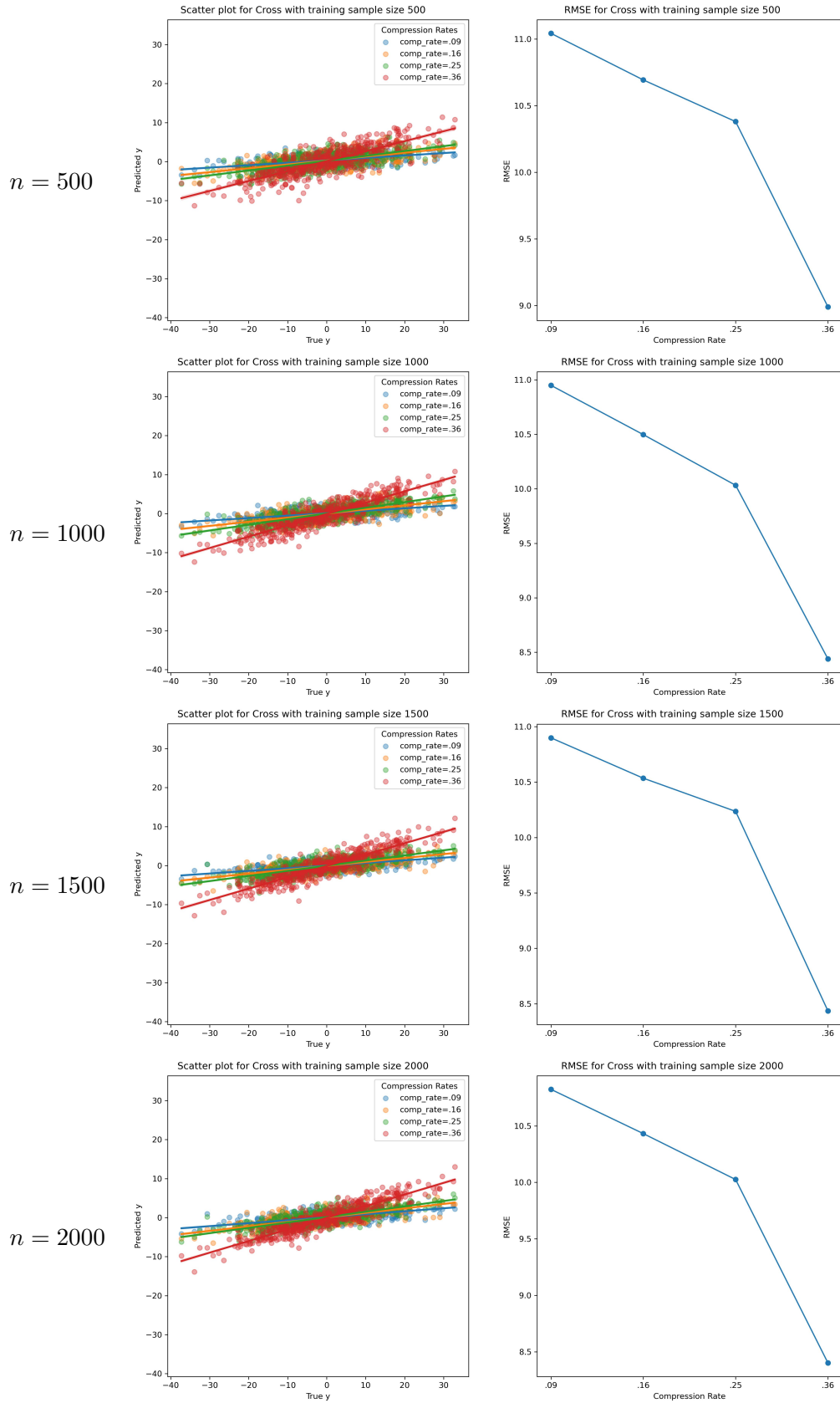


Figure 6: Prediction performances of different compression rates $r \in \{0.09, 0.16, 0.25, 0.36\}$ using different training sample size $n \in \{500, 1000, 1500, 2000\}$ (rows). Left column: scatter plots of actual data (horizontal axis) versus predicted data (vertical axis) with regression lines for different compression rates in different colors ($r = 0.09$: blue, $r = 0.16$: orange, $r = 0.25$: green, and $r = 0.36$: red). Right column: prediction RMSE (vertical axis) for different compression rates (horizontal axis).

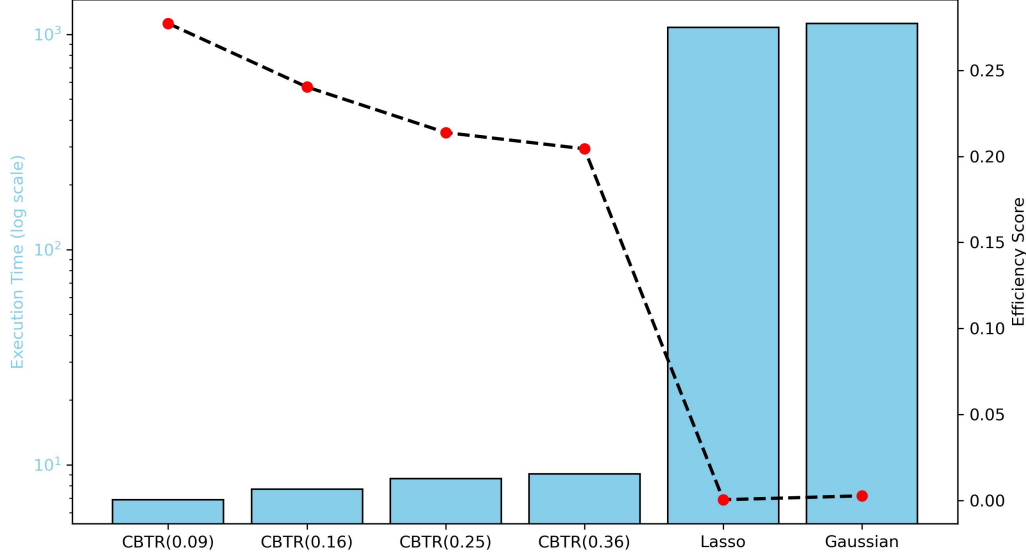


Figure 7: Total computational cost in a log scale (blue bars, left vertical axis) and efficiency scores (red dots, right vertical axis) for the Compressed Bayesian Tensor Regression with different compression rates $r \in \{0.09, 0.16, 0.25, 0.36\}$ (CBTR(r)), the Bayesian Lasso regression (Lasso) and the Gaussian regression (Gaussian).

Fig. 6 shows the prediction performances (out-of-sample scatter plots and RMSE plots) with different compression rates and different training sample sizes for the setting ‘Cross’. The random projection is performed with the first mode preserved. From the scatter plots, it’s clear that as the compression rate increases, the slope of the regression line of the scatter points increases, suggesting a better prediction performance. This is also shown in the RMSE plots in the right column of Fig. 6.

The computational cost of CBTR with different compression rates ($r \in \{0.09, 0.16, 0.25, 0.36\}$) is compared to that of Bayesian tensor regression using Gaussian priors and Lasso priors. The computational time is obtained for the simulation setting **CR** with a tensor coefficient size of 60×60 and a training sample size of $n = 2000$. 1000 Gibbs iterations are used to sample the unknowns. The left axis of Fig. 7 shows the computational time in a log scale. As the compression rate increases, the computational time increases; however, compared to BTR with Lasso and Gaussian priors, CBTR is faster by an order of 2. To measure the performance-per-cost of different models, we report the efficiency scores, which are computed as follows:

$$\text{Efficiency Score} = \frac{1}{\text{RMSE} \times \text{Cost}} \quad (20)$$

where “Cost” is the computational cost measured in hours. A higher efficiency score suggests better performance-per-cost. The black dashed line with red dots in Fig. 7 shows the efficiency scores of different models. CBTR had much higher efficiency scores compared to BTR with Lasso and Gaussian priors. Not surprisingly, as the compression rate increases, the efficiency scores decrease.

5.2 Empirical application

We demonstrate the performance of compressed Bayesian tensor regression (CBTR) using a real-world application studying the effects of oil volatility on the return of stock markets (S&P 500). We apply our tensor regression framework to a large dataset with mixed-frequency variables, as used in Casarin et al. (2025). We regress monthly log-returns of S&P500 (SP) on covariates sampled at daily frequency with monthly lags ranging from one to four. The

daily observations that we included are good oil volatility (GV), bad oil volatility (BV), US dollar index (ER), TED spread (IR), VIX index (VI), T-bill rate (TB), and bond spread (BD). Thus, the tensor predictors and coefficients are of size $(4, 7, 22)$, which corresponds to the number of temporal lags, number of regressors, and number of daily observations per month. We use 350 observations as training samples and 31 observations as testing samples. A representation of the model is:

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

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_3 th slice of tensor coefficients B along the third mode. The conditional mean of the model in (21) is given as the sum over slices corresponding to different temporal lags (third mode).

In Fig. D.1 of Appendix D, we compare the in-sample fittings as well as out-of-sample predictions of tensor regression without applying random projection and with different random projection methods (TW: tensor-wise without mode preservation, MW: mode-wise without mode preservation, MW(1): mode-wise preserving first mode, MW(1,2): mode-wise preserving first and second mode). As shown in the figure, the in-sample fittings of BTR and CBTR are relatively similar. This is also reflected in the RMSE reported in Table 1.

Table 1: Root Mean Square (Forecasting) Errors for in-sample fitting (out-of-sample forecasting) of Bayesian Tensor Regression (BTR) and Compressed Bayesian Tensor Regressions (CBTR) with different random projection types.

	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-of-sample	0.1148	0.0676	0.0623	0.0723	0.0383	0.0600	0.0508

However, the credible interval of the BTR appears to cover the actual data more effectively than that of the CBTR. More importantly, what differentiates CBTR from Bayesian tensor regression (BTR) is its out-of-sample forecasting abilities. Where every different random projection method outperforms BTR. Between different CBTRs, MW performs better than TW in terms of RMSE, which coincides with the simulation results. Between MW models, those preserving modes (MW(1) and MW(1,2)) perform better than those not preserving modes (MW). The performances of preserving 1 and 2 modes are very close, where preserving 2 modes offers slightly better in-sample fitting but worse out-of-sample forecasting.

The empirical application demonstrates the validity of random projection in reducing data dimensionality while preserving important information for making inferences and forecasting. The fact that CBTR outperforms BTR in forecasting is encouraging. Moreover, we explore different types of random projection methods and find out that CBTR-MW performs better than CBTR-TW in both simulation and empirical applications.

6 Conclusion

This paper introduces a Compressed Bayesian Tensor Regression (CBTR) framework that efficiently addresses the challenges of high-dimensional tensor covariates through a novel Generalized Tensor Random Projection (GTRP) strategy. The proposed method extends existing tensor projection approaches by allowing both mode-wise and tensor-wise projections, offering flexibility to preserve or reduce tensor modes and dimensions. Theoretical guarantees are

provided in the form of concentration inequalities and posterior consistency results, ensuring that inference and prediction remain valid after compression.

We design a Gibbs sampling algorithm tailored to hierarchical priors, including PARAFAC-based shrinkage priors, and introduce Bayesian model averaging to account for variability introduced by random projections. Our extensive simulation studies demonstrate that CBTR achieves substantial computational gains and improved prediction accuracy compared to standard Bayesian tensor regression, especially when the random projection preserves meaningful tensor structures. These findings are reinforced by an empirical application to financial data, where CBTR outperforms its uncompressed counterpart in out-of-sample forecasting.

Overall, our work establishes CBTR as a scalable and theoretically grounded alternative to conventional tensor regression methods, with potential for application in a wide range of domains involving structured, high-dimensional data. Future research will explore extensions to Kronecker-based projections, non-Gaussian likelihoods, and dependent data structures, opening further opportunities for efficient Bayesian learning in complex environments.

A Proofs of the results

A.1 Proof of Proposition 1

When $R = 0$ and $M = 1$ the projection writes as a scalar product between vector and a matrix, that is $\text{GTRP}(\mathcal{X}_j) = \mathcal{X}_j \times_{1:N} \mathcal{H}_{1:N} = \sum_{j_1=1}^{p_1} \cdots \sum_{j_N=1}^{p_N} \mathcal{X}_{j,j_1,\dots,j_N} \mathcal{H}_{j_1,\dots,j_N,:} = \text{vec}(\mathcal{X}_j) \text{mat}_{1:N}(\mathcal{H})$ where $\mathcal{H}$ is a $N+1$ -mode projection tensor with iid entries. $\text{vec}(\cdot)$ is a vectorization operator and $\text{mat}_{1:N}(\cdot)$ is a matricisation operator stacking in one mode all elements from mode 1 to mode N (e.g., see Hackbusch, 2019, Ch. 5). The proof follows by setting $d = p_1 \cdots p_N$ and $k = q_1$ in JL's Lemma of Achlioptas (2003).

A.2 Proof of Theorem 1

Before proving the theorem, we provide some preliminary results.

Lemma 1. *Let $\mathcal{T} = \tau_1 \otimes \cdots \otimes \tau_N$ be a $q_1 \times \cdots \times q_N$ tensor with $\tau_m = \mathbf{t}' T_m / \sqrt{p_m}$, where T_m are independent normal $p_m \times q_m$ projection matrices such that $T_{j_m, i_m} \sim \mathcal{N}(0, 1/p_m)$. With entries of $\mathcal{T}$ are $\mathcal{T}_{i_1, \dots, i_N} = \tau_{1, i_1} \cdots \tau_{N, i_N}$ with $\tau_{m, i_m} \sim \mathcal{N}(0, 1/p_m)$ independent normal. Let*

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

be the $q_1 \times \cdots \times q_N$ tensor obtained by projecting the rescaled $p_1 \times \cdots \times p_N$ unit tensor. The tensor entries $\mathcal{Q}_{i_1, \dots, i_N}(\mathcal{A})$ of the tensor $\mathcal{Q}(\mathcal{A})$ satisfy the following properties

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

Proof. Without loss of generality, we prove the results for the case $N = 3$.

- i. This follows by the same argument as in the proof of Lemma 6.1 in Achlioptas (2003).
- ii.

$$\mathbb{E}(\mathcal{Q}_{i_1, i_2, i_3}^{2k}) = \mathbb{E} \left(\left(\frac{1}{\sqrt{p_1 p_2 p_3}} \sum_{j_1=1}^{p_1} \sum_{j_2=1}^{p_2} \sum_{j_3=1}^{p_3} H_{1, j_1, i_1} H_{2, j_2, i_2} H_{3, j_3, i_3} \right)^{2k} \right) \quad (\text{A.2})$$

$$= \mathbb{E} \left(\left(\frac{1}{\sqrt{p_1 p_2 p_3}} \sum_{j_1=1}^{p_1} \sum_{j_2=1}^{p_2} H_{1, j_1, i_1} H_{2, j_2, i_2} \sum_{j_3=1}^{p_3} H_{3, j_3, i_3} \right)^{2k} \right) \quad (\text{A.3})$$

$$= \mathbb{E} \left(\left(\frac{1}{\sqrt{p_1 p_2}} \frac{H_{3,i_3}}{\sqrt{p_3}} \sum_{j_1=1}^{p_1} \sum_{j_2=1}^{p_2} H_{1,j_1,i_1} H_{2,j_2,i_2} \right)^{2k} \right) \quad (\text{A.4})$$

$$= \mathbb{E} \left(\left(\frac{H_{1,i_1}}{\sqrt{p_1}} \frac{H_{2,i_2}}{\sqrt{p_2}} \mathbf{h}_{3,i_3} \right)^{2k} \right) \quad (\text{A.5})$$

$$= \mathbb{E} \left((\mathbf{h}_{1,i_1})^{2k} \right) \mathbb{E} \left((\mathbf{h}_{2,i_2})^{2k} \right) \mathbb{E} \left((\mathbf{h}_{3,i_3})^{2k} \right) \quad (\text{A.6})$$

$$\leq \mathbb{E} \left((\boldsymbol{\tau}_{1,i_1})^{2k} \right) \mathbb{E} \left((\boldsymbol{\tau}_{2,i_2})^{2k} \right) \mathbb{E} \left((\boldsymbol{\tau}_{3,i_3})^{2k} \right) \quad (\text{A.7})$$

$$= \mathbb{E} \left((\boldsymbol{\tau}_{1,i_1} \boldsymbol{\tau}_{2,i_2} \boldsymbol{\tau}_{3,i_3})^{2k} \right) \quad (\text{A.8})$$

$$= \mathbb{E} \left((\mathcal{T}_{i_1,i_2,i_3})^{2k} \right) \quad (\text{A.9})$$

the inequality follows using the same argument as in (Achlioptas, 2003, Lemma 6.2). $\square$

Lemma 2. Let $x_j \stackrel{\text{ind}}{\sim} \mathcal{G}a(\alpha, \beta_j)$ with pdf

$$f(x) = \frac{\beta_j^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta_j x}, \quad x > 0$$

$\mathbb{E}(e^{hx_1 \cdots x_N}) = \left(\frac{1}{\Gamma(\alpha)} \right)^N G_{N,1}^{1,N} \left(-\frac{h}{\beta_1 \cdots \beta_N} \middle| \begin{matrix} 1-\alpha, \dots, 1-\alpha \\ 0 \end{matrix} \right)$, where $G_{p,q}^{m,n}(\cdot | \begin{smallmatrix} a_1, \dots, a_p \\ b_1, \dots, b_q \end{smallmatrix})$ is the Meijer G-function given in (Mathai et al., 2010, Def. 1.5).

Proof. Let $H_{p,q}^{m,n} \left(\cdot \middle| \begin{smallmatrix} (a_1, A_1), \dots, (a_p, A_p) \\ (b_1, B_1), \dots, (b_q, B_q) \end{smallmatrix} \right)$ be the Fox H-function given in (Mathai et al., 2010, Def. 1.1) and define $z = x_2 \cdots x_N$. Since $\exp\{hxz\} = H_{0,1}^{1,0} \left(-hzx \middle| \begin{smallmatrix} - \\ (0, 1) \end{smallmatrix} \right)$ (Mathai et al., 2010, Eq. 1.39), then by the law of iterated expectation

$$\begin{aligned} & \mathbb{E} \left(\mathbb{E} \left(e^{hx_1 z} \middle| z \right) \right) \\ &= \mathbb{E} \left(\frac{\beta_1^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\beta_1 x} x^{\alpha-1} e^{hxz} dx \right) \\ &= \mathbb{E} \left(\frac{\beta_1^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\beta_1 x} x^{\alpha-1} H_{0,1}^{1,0} \left(-hzx \middle| \begin{smallmatrix} - \\ (0, 1) \end{smallmatrix} \right) dx \right) \end{aligned} \quad (\text{A.10})$$

which is the Laplace transform of $x^{\alpha-1} H_{0,1}^{1,0} \left(-hzx \middle| \begin{smallmatrix} - \\ (0, 1) \end{smallmatrix} \right)$. From Eq. 2.19 in Mathai et al. (2010), with $\varrho = \alpha$, $a = -hz$ and $s = \beta$, Eq. A.10 becomes

$$\mathbb{E} \left(\frac{\beta_1^\alpha}{\Gamma(\alpha)} \beta_1^{-\alpha} H_{1,1}^{1,1} \left(-\frac{hz}{\beta_1} \middle| \begin{smallmatrix} (1-\alpha, 1) \\ (0, 1) \end{smallmatrix} \right) \right) = \dots \quad (\text{A.11})$$

$$= \mathbb{E} \left(\left(\frac{1}{\Gamma(\alpha)} \right)^{N-1} H_{N-1,1}^{1,N-1} \left(-\frac{h}{\beta_1 \cdots \beta_{N-1}} \middle| \begin{smallmatrix} (1-\alpha, 1), \dots, (1-\alpha, 1) \\ (0, 1) \end{smallmatrix} \right) \right) \quad (\text{A.12})$$

$$= \frac{1}{\Gamma(\alpha)^N} \beta_N^\alpha \int_0^\infty e^{-\beta_N x} x^{\alpha-1} H_{N-1,1}^{1,N-1} \left(-\frac{h}{\beta_1 \cdots \beta_{N-1}} \middle| \begin{smallmatrix} (1-\alpha, 1), \dots, (1-\alpha, 1) \\ (0, 1) \end{smallmatrix} \right) dx \quad (\text{A.13})$$

$$= \Gamma(\alpha)^{-N} H_{N,1}^{1,N} \left(-\frac{h}{\beta_1 \cdots \beta_N} \middle| \begin{smallmatrix} (1-\alpha, 1), \dots, (1-\alpha, 1) \\ (0, 1) \end{smallmatrix} \right) \quad (\text{A.14})$$

$$= \Gamma(\alpha)^{-N} G_{N,1}^{1,N} \left(-\frac{h}{\beta_1 \cdots \beta_N} \middle| \begin{smallmatrix} 1-\alpha, \dots, 1-\alpha \\ 0 \end{smallmatrix} \right) \quad (\text{A.15})$$

$$= \Gamma(\alpha)^{-N} G_{1,N}^{N,1} \left(-\frac{\beta_1 \cdots \beta_N}{h} \middle| \begin{matrix} 1 \\ 1-\alpha, \dots, 1-\alpha \end{matrix} \right) \quad (\text{A.16})$$

where the before last equality follows from the definition of Meijer G-function $G_{p,q}^{m,n}(\cdot | \begin{smallmatrix} a_1, \dots, a_p \\ b_1, \dots, b_q \end{smallmatrix})$ given in (Mathai et al., 2010, Def. 1.5), and the last equality from Eq. (1.58) in Mathai et al. (2010). $\square$

Lemma 3.

$$\mathbb{E}(\exp\{h\mathcal{Q}_{1,\dots,1}(\mathcal{A})^2\}) \leq \frac{1}{\pi^{N/2}} G_{1,N}^{N,1} \left(\frac{1}{p(N)2^N h} \middle| \begin{matrix} 1 \\ 1/2, \dots, 1/2 \end{matrix} \right) \quad (\text{A.17})$$

Proof. By Monotone Convergence Theorem

$$\mathbb{E}(\exp\{h\mathcal{Q}_{1,\dots,1}(\mathcal{A})^2\}) = \sum_{k=0}^{\infty} \frac{h^k}{k!} \mathbb{E}(\mathcal{Q}_{1,\dots,1}(\mathcal{A})^{2k}) \quad (\text{A.18})$$

$$\leq \sum_{k=0}^{\infty} \frac{h^k}{k!} \mathbb{E}(\mathcal{T}_{1,\dots,1}^{2k}) = \mathbb{E}(\exp\{h\mathcal{T}_{1,\dots,1}^2\}) \quad (\text{A.19})$$

$$= \frac{1}{\pi^{N/2}} G_{1,N}^{N,1} \left(-\frac{p(N)}{2^N h} \middle| \begin{matrix} 1 \\ 1/2, \dots, 1/2 \end{matrix} \right) \quad (\text{A.20})$$

where the inequality follows from Lemma 1 and the last equality from Lemma 2, where we set $\alpha = 1/2$ and $\beta_j = p_j/2$ in the Meijer G-function, and from $\Gamma(1/2) = \sqrt{\pi}$. $\square$

A.2.1 Proof of Theorem 1

The $(i_1, \dots, i_N)$ -th element of $f(\mathcal{U})$ write as:

$$f(\mathcal{U})_{i_1, \dots, i_N} = \sqrt{C(N, M)} \sum_{j_1=1}^{p_1} \cdots \sum_{j_N=1}^{p_N} \mathcal{X}_{t, j_1, \dots, j_N} H_{1, j_1, i_1} \cdots H_{N, j_N, i_N} \quad (\text{A.21})$$

We denote with $\|f(\mathcal{U})\|$ the Frobenius' norm of $f(\mathcal{U})$ and prove that

$$\|\mathcal{U} - \mathcal{V}\|^2(1 - \varepsilon) \leq \|f(\mathcal{U}) - f(\mathcal{V})\|^2 \leq \|\mathcal{U} - \mathcal{V}\|^2(1 + \varepsilon) \quad (\text{A.22})$$

with probability at least $1 - \kappa_n$ for any pair $\mathcal{U}, \mathcal{V} \in \mathbb{R}^{p_1 \times \dots \times p_N}$.

Since the map satisfies $f(\mathcal{U}) - f(\mathcal{V}) = f(\mathcal{U} - \mathcal{V})$ the statement becomes

$$\|\mathcal{A}\|^2(1 - \varepsilon) \leq \|f(\mathcal{A})\|^2 \leq \|\mathcal{A}\|^2(1 + \varepsilon) \quad (\text{A.23})$$

with probability at least $1 - \kappa_n$. Since $\|f(\mathcal{A})\|^2$ is proportional to $\|\mathcal{A}\|^2$ it is sufficient to prove the following

$$(1 - \varepsilon) \leq \|f(\mathcal{A})\|^2 \leq (1 + \varepsilon) \quad (\text{A.24})$$

Define $S(\mathcal{A}) = \|f(\mathcal{A})\|^2 / C(N, M)$ and $\mathcal{Q}(\mathcal{A})$ as the tensor with elements

$$\mathcal{Q}_{i_1, \dots, i_N}(\mathcal{A}) = \sum_{j_1=1}^{p_1} \cdots \sum_{j_N=1}^{p_N} \mathcal{A}_{j_1, \dots, j_N} H_{1, j_1, i_1} \cdots H_{N, j_N, i_N} \quad (\text{A.25})$$

Then $S(\mathcal{A}) = \sum_{i_1=1}^{q_1} \cdots \sum_{i_N=1}^{q_N} \mathcal{Q}_{i_1, \dots, i_N}(\mathcal{A})^2$. By Markov's inequality, it follows

$$P(\{\|f(\mathcal{A})\|^2 > (1 + \varepsilon)\}) = P\left(\left\{\exp\{hS(\mathcal{A})\} > \exp\left\{\frac{h}{C(N, M)}(1 + \varepsilon)\right\}\right\}\right) \quad (\text{A.26})$$

$$\leq \mathbb{E}(\exp\{hS(\mathcal{A})\}) \exp\left\{-\frac{h}{C(N, M)}(1 + \varepsilon)\right\} \quad (\text{A.27})$$

$$\leq (\mathbb{E}(\exp\{h\mathcal{Q}_{1,\dots,1}(\mathcal{A})^2\}))^{q(N)} \exp\left\{-\frac{h}{C(N,M)}(1+\varepsilon)\right\} \quad (\text{A.28})$$

$$\leq \left(f(h) \exp\left\{-\frac{h}{p(N)}(1+\varepsilon)\right\}\right)^{q(N)} \quad (\text{A.29})$$

The last inequality follows from Lemma 3, where we defined

$$f(h) = \frac{1}{\pi^{N/2}} G_{1,N}^{N,1} \left(-\frac{p(N)}{2^N h} \middle| \begin{matrix} 1 \\ 1/2, \dots, 1/2 \end{matrix} \right) \quad (\text{A.30})$$

One can obtain the optimal exponential bound for the upper tail by optimizing in h . However, the first-order condition is intractable due to the presence of the Meijer G-function. But a “good enough” solution of h can be obtained using power-log expansion for the Meijer G-function as in Stojanac et al. (2018).

The first order condition of (A.29) with respect to h after simplification is

$$\frac{1}{h} G_{N,1}^{1,N} \left(\frac{2^N h}{p(N)} \middle| \begin{matrix} 1/2, \dots, 1/2 \\ 1 \end{matrix} \right) + \frac{1+\epsilon}{p(N)} G_{N,1}^{1,N} \left(\frac{2^N h}{p(N)} \middle| \begin{matrix} 1/2, \dots, 1/2 \\ 0 \end{matrix} \right) = 0 \quad (\text{A.31})$$

Applying the lowest order power-log series expansion for the above Meijer G-function

$$G_{1,N}^{N,1} \left(\frac{2^N h}{p(N)} \middle| \begin{matrix} 1/2, \dots, 1/2 \\ x \end{matrix} \right) \approx \left(\frac{2^N h}{p(N)} \right)^{\frac{1}{2}} \bar{H}_{0,N-1}^x \left[\log\left(\frac{2^N h}{p(N)}\right) \right]^{N-1} \quad (\text{A.32})$$

where

$$\begin{aligned} \bar{H}_{0,N-1}^0 &= -\frac{1}{(N-1)!} \Gamma\left(\frac{1}{2}\right) \\ \bar{H}_{0,N-1}^1 &= \frac{1}{2} \bar{H}_{0,N-1}^0 \end{aligned}$$

Equation (A.31) approximates as follows

$$\frac{1}{h} \left(\frac{2^N h}{p(N)} \right)^{\frac{1}{2}} \frac{1}{2} \bar{H}_{0,N-1}^0 \left[\log\left(\frac{2^N h}{p(N)}\right) \right]^{N-1} + \quad (\text{A.33})$$

$$\frac{1+\epsilon}{p(N)} \left(\frac{2^N h}{p(N)} \right)^{\frac{1}{2}} \bar{H}_{0,N-1}^0 \left[\log\left(\frac{2^N h}{p(N)}\right) \right]^{N-1} = 0 \quad (\text{A.34})$$

$$\left(\frac{2^N h}{p(N)} \right)^{\frac{1}{2}} \bar{H}_{0,N-1}^0 \left[\log\left(\frac{2^N h}{p(N)}\right) \right]^{N-1} \left(\frac{1}{2h} + \frac{1+\epsilon}{p(N)} \right) = 0 \quad (\text{A.35})$$

Since $h > 0$, the only solution is $h = p(N)/2^N$, and it follows that

$$\begin{aligned} &P(\{\|f(\mathcal{A})\|^2 > (1+\varepsilon)\}) \\ &\leq \left(\frac{1}{\pi^{N/2}} G_{1,N}^{N,1} \left(1 \middle| \begin{matrix} 1 \\ 1/2, \dots, 1/2 \end{matrix} \right) \exp\left\{-\frac{1}{2^N}(1+\epsilon)\right\} \right)^{q(N)} \\ &= \exp\left\{q(N) \left(-\frac{N}{2} \ln \pi + \ln G_{1,N}^{N,1} \left(1 \middle| \begin{matrix} 1 \\ 1/2, \dots, 1/2 \end{matrix} \right) - \frac{1+\epsilon}{2^N} \right)\right\} \end{aligned} \quad (\text{A.36})$$

Given that the Meijer G-function is fully specified, we can evaluate its value and the above bound can be approximated as

$$\begin{aligned} &P(\{\|f(\mathcal{A})\|^2 > (1+\varepsilon)\}) \\ &\leq \exp\left\{q_1 q_2 q_3 \left(-\frac{7}{100} - \frac{1+\epsilon}{8} \right)\right\} \end{aligned} \quad (\text{A.37})$$

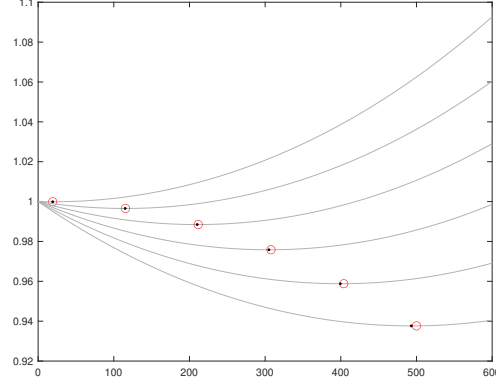


Figure A.1: The plot of the lower bound as a function of h for different values of $\epsilon = 0.01, 0.06, 0.11, 0.16, 0.21, 0.26$. The optimal values of h that minimize the bound are shown in the black dots, the approximated values of h are shown as red circles.

For the lower tail exponential bound, consider

$$P(\{\|f(\mathcal{A})\|^2 < (1 - \epsilon)\}) = P\left(\{\exp\{-hS(\mathcal{A})\} > \exp\left\{-\frac{h}{C(N, M)}(1 - \epsilon)\right\}\right) \quad (\text{A.38})$$

$$\leq \mathbb{E}(\exp\{-hS(\mathcal{A})\}) \exp\left\{\frac{h}{C(N, M)}(1 - \epsilon)\right\} \quad (\text{A.39})$$

$$\leq (\mathbb{E}(\exp\{-h\mathcal{Q}_{1, \dots, 1}(\mathcal{A})^2\}))^{q(N)} \exp\left\{\frac{h}{C(N, M)}(1 - \epsilon)\right\} \quad (\text{A.40})$$

By expanding $\exp\{-h\mathcal{Q}_{1, \dots, 1}(\mathcal{A})^2\}$ we have

$$P(\{\|f(\mathcal{A})\|^2 < (1 - \epsilon)\}) \quad (\text{A.41})$$

$$\leq \left(1 - h\mathbb{E}(\mathcal{Q}_{1, \dots, 1}(\mathcal{A})^2) + \frac{h^2}{2}\mathbb{E}(\mathcal{Q}_{1, \dots, 1}(\mathcal{A})^4)\right)^{q(N)} \exp\left\{\frac{h}{C(N, M)}(1 - \epsilon)\right\} \quad (\text{A.42})$$

$$\leq \left(1 - \frac{h}{p(N)} + \frac{3^N h^2}{2(p(N))^2}\right)^{q(N)} \exp\left\{\frac{h}{C(N, M)}(1 - \epsilon)\right\} \quad (\text{A.43})$$

To optimize the bound, solving the first order condition of (A.43) with respect to h , this gives $h = \sqrt{\frac{2(p(N))^2\epsilon}{3^N(1-\epsilon)} + \left(\frac{(p(N))(3^N-1+\epsilon)}{3^N(1-\epsilon)}\right)^2} - \frac{(p(N))(3^N-1+\epsilon)}{3^N(1-\epsilon)}$. Numerical studies (Figure A.1) suggest $h = \frac{p(N)}{3^N-1}\epsilon$ is a good approximation. Substituting this value of h , we get (A.45), series expansion gives (A.46).

$$P(\{\|f(\mathcal{A})\|^2 < (1 - \epsilon)\}) \quad (\text{A.44})$$

$$< \exp\left\{(q(N)) \ln\left(1 - \frac{\epsilon}{3^N-1} + \frac{3^N \epsilon^2}{2(3^N-1)^2}\right) + \frac{q(N)}{3^N-1}\epsilon(1 - \epsilon)\right\} \quad (\text{A.45})$$

$$\approx \exp\left\{-q(N) \left(\frac{\epsilon^2}{2(3^N-1)} - \frac{(3^{N+1}-2)\epsilon^3}{6(3^N-1)^3}\right)\right\} \quad (\text{A.46})$$

To get JL-embedding, we need $2 \times \exp\left\{-q(N) \left(\frac{\epsilon^2}{2(3^N-1)} - \frac{(3^{N+1}-2)\epsilon^3}{6(3^N-1)^3}\right)\right\} \leq \frac{2}{n^{2+\beta}}$, thus $q(N) \geq \frac{4+2\beta}{\frac{\epsilon^2}{3^N-1} - \frac{(3^{N+1}-2)\epsilon^3}{3(3^N-1)^3}} \log n$.

A.3 Proof of Theorem 2

Note that the $(i_1, i_2, \dots, i_N)$ -th entry from our mode-wise random projection can be written equivalently as the inner product of the tensor $\mathcal{X}$ and a rank 1 tensor constructed by the outer product of the corresponding columns of matrices $H_{n, :, i_n}$:

$$f(\mathcal{X})_{i_1, \dots, i_N} = \frac{1}{\sqrt{q(N)}} \langle H_{1, :, i_1} \circ H_{2, :, i_2} \circ \dots \circ H_{N, :, i_N}, \mathcal{X} \rangle = \frac{1}{\sqrt{q(N)}} u_{i_1, \dots, i_N}$$

To find the bound on the embedding dimensions, we follow the similar arguments from Rakhshan and Rabusseau (2020) to first bound the variance of the Frobenius norm of $f(\mathcal{X})$ and then applying Hypercontractivity Concentration Inequality (Schudy and Sviridenko, 2012) to bound the embedding dimension.

$$\mathbb{V}(\|f(\mathcal{X})\|_F^2) = \mathbb{E}\|f(\mathcal{X})\|_F^4 - (\mathbb{E}\|f(\mathcal{X})\|_F^2)^2$$

Due to expected isometry, it can be shown that $\mathbb{E}\|f(\mathcal{X})\|_F^2 = \|\mathcal{X}\|_F^2 = 1$, and

$$\mathbb{E}\|\mathcal{U}\|_F^4 = \sum_{i_1=1}^{q_1} \dots \sum_{i_N=1}^{q_N} \mathbb{E}u_{i_1, \dots, i_N}^4 + \sum_{i_1 \dots i_N \neq i'_1 \dots i'_N} \mathbb{E}(u_{i_1, \dots, i_N}^2 u_{i'_1, \dots, i'_N}^2)$$

Since $u_{i_1, \dots, i_N}^2$ and $u_{i'_1, \dots, i'_N}^2$ are independent, the second term on the right hand side amounts to $q(N)(q(N) - 1)\|\mathcal{X}\|_F^4 = q(N)(q(N) - 1)$. Using the same argument in Rakhshan and Rabusseau (2020) we can bound $\mathbb{E}u_{i_1, \dots, i_N}^4$,

$$\begin{aligned} \mathbb{E}u_{i_1, \dots, i_N}^4 &= \mathbb{E} \langle H_{1, :, i_1} \circ H_{2, :, i_2} \circ \dots \circ H_{N, :, i_N}, \mathcal{X} \rangle^4 \\ &\leq 3^N \|f(\mathcal{X})\|_F^4 \\ &= 3^N \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{V}(\|f(\mathcal{X})\|_F^2) &= \mathbb{V}\left(\left\|\frac{1}{\sqrt{q(N)}}\mathcal{U}\right\|_F^2\right) \\ &= \frac{1}{q(N)^2} \left(\mathbb{E}\|\mathcal{U}\|_F^4 - (\mathbb{E}\|\mathcal{U}\|_F^2)^2\right) \\ &\leq \frac{1}{q(N)^2} [q(N)3^N + q(N)(q(N) - 1)] - 1 \\ &= \frac{3^N - 1}{q(N)} \end{aligned}$$

By Hypercontractivity Concentration Inequality, for some positive constants C and K we have,

$$\begin{aligned} \mathbb{P}(\left|\|f(\mathcal{X})\|_F^2 - \|\mathcal{X}\|_F^2\right| \geq \varepsilon \|\mathcal{X}\|_F^2) &\leq C \exp \left[- \left(\frac{\varepsilon^2}{K \mathbb{V}(\|f(\mathcal{X})\|_F^2)} \right)^{\frac{1}{2N}} \right] \\ &\leq C \exp \left[- \frac{(\sqrt{q(N)}\varepsilon)^{\frac{1}{N}}}{(K 3^N)^{\frac{1}{2N}}} \right] \end{aligned}$$

A.4 Proof of Theorem 3

Let $\mathcal{P}_n$ denote a sequence of sets of probability densities, $N(\varepsilon_n, \mathcal{P}_n)$ the minimum number of Hellinger balls of radius ε_n needed to cover $\mathcal{P}_n$. Define the following conditions:

- a) $\log N(\varepsilon_n, \mathcal{P}_n) \leq n\varepsilon_n^2$ for all large n
- b) $\pi(\mathcal{P}_n^c) \leq e^{-2n\varepsilon_n^2}$ for all large n
- c) $\pi \left[f : d_t(f, f_0) < \frac{\varepsilon_n^2}{4} \right] \geq e^{-n\varepsilon_n^2/4}$ for all large n .

Proposition 2. *If $n\varepsilon_n^2 \rightarrow \infty$, then under conditions a, b, c (for some $t > 0$), we have*

$$E_{f_0} \pi [d(f, f_0) > 4\varepsilon_n \mid (y_i, \mathcal{X}_i)_{i=1}^n] \leq 4e^{-n\varepsilon_n^2 \min(1/2, t/4)}$$

Proposition 2 has been proved in Jiang (2007). We prove Theorem 3 by showing conditions a, b and c hold in our case for some positive t .

Proposition 3. *Assume $\mathcal{B} \sim \mathcal{TN}(\mathbf{0}, \Sigma_1, \dots, \Sigma_N)$, where $\mathcal{TN}$ denotes the Tensor Normal distribution and Σ_n is the covariance matrix for mode n . Then*

$$P(|\langle f(\mathcal{X}), \mathcal{B} \rangle - \langle \mathcal{X}, \mathcal{B}_0 \rangle| < \Delta) > P(X - Y \geq 2),$$

where $X \sim \text{Poi}(\frac{\Delta_1}{2})$, $Y \sim \text{Poi}(\frac{\lambda}{2})$ with $\Delta_1 = \frac{\Delta^2}{\text{Var}(\langle f(\mathcal{X}), \mathcal{B} \rangle)}$, $\lambda = \frac{\langle \mathcal{X}, \mathcal{B}_0 \rangle^2}{\text{Var}(\langle f(\mathcal{X}), \mathcal{B} \rangle)}$, $\text{Var}(\langle f(\mathcal{X}), \mathcal{B} \rangle) = \text{vec}(f(\mathcal{X}))'(\Sigma_1 \otimes \dots \otimes \Sigma_N)\text{vec}(f(\mathcal{X}))$.

Proof. Note that $\langle f(\mathcal{X}), \mathcal{B} \rangle \sim \mathcal{N}(0, \text{vec}(f(\mathcal{X}))'(\Sigma_1 \otimes \dots \otimes \Sigma_N)\text{vec}(f(\mathcal{X})))$. This implies

$$\frac{|\langle f(\mathcal{X}), \mathcal{B} \rangle - \langle \mathcal{X}, \mathcal{B}_0 \rangle|^2}{\text{Var}(\langle f(\mathcal{X}), \mathcal{B} \rangle)} \sim \chi_1^2(\lambda),$$

where $\chi_1^2(\lambda)$ is the noncentral chi-squared distribution with degrees of freedom 1 and noncentral parameter $|\lambda|\sqrt{\text{Var}(\langle f(\mathcal{X}), \mathcal{B} \rangle)} = |\mathbb{E}(\langle f(\mathcal{X}), \mathcal{B} \rangle - \langle \mathcal{X}, \mathcal{B}_0 \rangle)| = \langle \mathcal{X}, \mathcal{B}_0 \rangle$. It is known that a noncentral chi-squared distribution can also be written a gamma mixture with Poisson weights

$$\begin{aligned} P(|\langle f(\mathcal{X}), \mathcal{B} \rangle - \langle \mathcal{X}, \mathcal{B}_0 \rangle| < \Delta) &= P\left(\frac{|\langle f(\mathcal{X}), \mathcal{B} \rangle - \langle \mathcal{X}, \mathcal{B}_0 \rangle|^2}{\text{Var}(\langle f(\mathcal{X}), \mathcal{B} \rangle)} < \Delta_1\right) \\ &= \sum_{i=0}^{\infty} \frac{e^{-\frac{\lambda}{2}} (\frac{\lambda}{2})^i}{i!} P(Z_{1+2i} < \Delta_1), \end{aligned} \quad (\text{A.47})$$

where $\Delta_1 = \Delta^2 / \text{Var}(\langle f(\mathcal{X}), \mathcal{B} \rangle)$ and $Z_{1+2i} \sim \chi_{1+2i}^2$. Note that $P(Z_{1+2i} < \Delta_1) > P(Z_{2+2i} < \Delta_1) = P(G < \Delta_1)$ where $G \sim \mathcal{Ga}(1+i, \frac{1}{2})$. From Proposition A.2 in Guhaniyogi and Dunson (2015), we obtain $P(|\langle f(\mathcal{X}), \mathcal{B} \rangle - \langle \mathcal{X}, \mathcal{B}_0 \rangle| < \Delta) > P(X - Y \geq 2)$. $\square$

Proof of Theorem 3. We will check the three conditions with $t = 1$. Let $b_n = \sqrt{8\tilde{\lambda}_n n\varepsilon_n^2}$.

Condition a. Let $\mathcal{P}_n$ be the set of all densities that can be represented by the $\mathcal{B}$ with entries $|b_{jkl}| < b_n$, $j = 1, \dots, q_{1,n}$, $k = 1, \dots, q_{2,n}$, $l = 1, \dots, q_{3,n}$. Let's consider the l^∞ balls of the form $(a_{jkl} - \delta, a_{jkl} + \delta)$ for each entry of the tensor coefficients with the center of each ball inside $\mathcal{P}_n$, the external covering number of $\mathcal{P}_n$ is bounded above by $(\frac{b_n}{\delta} + 1)^{q_n}$ where $q_n = q_{1,n}q_{2,n}q_{3,n}$.

Let f_u be any density in $\mathcal{P}_n$, $\exists \mathcal{B}$ s.t. $u = \langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle$, $|b_{jkl}| \leq b_n$ and

$$f_u(y) = \exp\{ya(u) + b(u) + c(y)\}$$

Let $b_{jkl} \in (c_{jkl} - \delta, c_{jkl} + \delta)$, s.t. $|b_{jkl} - c_{jkl}| \leq \delta$ and $|c_{jkl}| \leq b_n$. Let $v = \langle \text{GTRP}(\mathcal{X}_i), \mathcal{C} \rangle$, and

$$f_v(y) = \exp \{ya(v) + b(v) + c(y)\}$$

We then find the number of Hellinger balls that required to cover $\mathcal{P}_n$ by using the fact $d(f, f_0) \leq (d_{KL}(f, f_0))^{1/2}$, where

$$\begin{aligned} d_{KL}(f_u, f_v) &= \iint f_v \log \left(\frac{f_v}{f_u} \right) \nu_y(dy) \nu_{\mathcal{X}}(d\mathcal{X}) \\ &= \iint [y(a(v) - a(u)) + (b(v) - b(u))] f_v \nu_y(dy) \nu_{\mathcal{X}}(d\mathcal{X}) \\ &= \int [(a(v) - a(u)) E[y | \mathcal{X}] + (b(v) - b(u))] \nu_{\mathcal{X}}(d\mathcal{X}) \\ &= \int (v - u) \left[a'(u_v) \left(-\frac{b'(v)}{a'(v)} \right) + b'(u_v) \right] \nu_{\mathcal{X}}(d\mathcal{X}). \end{aligned} \quad (\text{A.48})$$

The last two steps are achieved by first integrating with respect to y and then applying the mean value theorem, where u_v is the intermediate point between u and v . By Cauchy-Schwartz inequality, the condition $|b_{jkl} - c_{jkl}| < \delta$ and the assumption $|x_{jkl}| < 1$ we have,

$$\begin{aligned} |u - v| &= |\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \text{GTRP}(\mathcal{X}_i), \mathcal{C} \rangle| = |\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} - \mathcal{C} \rangle| \\ &\leq \|\text{GTRP}(\mathcal{X}_i)\| \|\mathcal{B} - \mathcal{C}\| \leq \|\mathcal{X}_i\| \sqrt{q_n} \delta \leq \sqrt{p_n q_n} \delta = \theta_n \delta, \end{aligned}$$

where we defined $\theta_n = \sqrt{q_n p_n}$. Since $|u| = |\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle| \leq \|\text{GTRP}(\mathcal{X}_i)\| \|\mathcal{B}\| \leq \sqrt{q_n p_n} b_n \leq b_n \theta_n$, similarly, $|v| \leq b_n \theta_n$, thus $|u_v| \leq b_n \theta_n$. Combining the results and (A.48), we have,

$$\begin{aligned} d(f_u, f_v) &\leq \sqrt{d_{KL}(f_u, f_v)} \\ &\leq \sqrt{\int |v - u| \left| a'(u_v) \left(-\frac{b'(v)}{a'(v)} \right) + b'(u_v) \right| \nu_{\mathcal{X}}(d\mathcal{X})} \\ &\leq \sqrt{2\theta_n \delta \sup_{|h| \leq b_n \theta_n} |a'(h)| \sup_{|h| \leq b_n \theta_n} \left| \frac{b'(h)}{a'(h)} \right| \int \nu_{\mathcal{X}}(d\mathcal{X})}. \end{aligned}$$

Let $\delta = \varepsilon_n^2 / (2\theta_n \sup_{|h| \leq b_n \theta_n} |a'(h)| \sup_{|h| \leq b_n \theta_n} \left| \frac{b'(h)}{a'(h)} \right|)$, one gets $d(f_u, f_v) \leq \varepsilon_n$. The entropy of $\mathcal{P}_n$ is therefore bounded from above by

$$\begin{aligned} \left(1 + \frac{2b_n \theta_n}{\varepsilon_n^2} \sup_{|h| \leq b_n \theta_n} |a'(h)| \sup_{|h| \leq b_n \theta_n} \left| \frac{b'(h)}{a'(h)} \right| \right)^{q_n} &= \left(1 - \frac{1}{\varepsilon_n^2} + \frac{D(b_n \theta_n)}{\varepsilon_n^2} \right)^{q_n} \\ &\leq \left(\frac{D(b_n \theta_n)}{\varepsilon_n^2} \right)^{q_n} \end{aligned}$$

where we defined $D(R) = 1 + R \sup_{|h| \leq R} |a'(h)| \sup_{|h| \leq R} \left| \frac{b'(h)}{a'(h)} \right|$ and the inequality follows from the assumption $\varepsilon_n^2 < 1$. Thus the Hellinger covering number satisfied $N(\varepsilon_n, \mathcal{P}_n) \leq \left(\frac{D(b_n \theta_n)}{\varepsilon_n^2} \right)^{q_n}$, implying $\log N(\varepsilon_n, \mathcal{P}_n) \leq q_n (\log D(b_n \theta_n) + \log(1/\varepsilon_n^2))$. Using the assumptions in *i*) of Theorem 3, $\frac{q_n \log(1/\varepsilon_n^2)}{n\varepsilon_n^2} \rightarrow 0$ and $\frac{q_n \log D(b_n \theta_n)}{n\varepsilon_n^2} \rightarrow 0$, condition *a* follows.

Condition b.

By union bound inequality, it follows:

$$\pi(\mathcal{P}_n^c) = \pi \left(\bigcup_{j=1}^{q_{1,n}} \bigcup_{k=1}^{q_{2,n}} \bigcup_{l=1}^{q_{3,n}} |b_{jkl}| > b_n \right) \leq \sum_{j=1}^{q_{1,n}} \sum_{k=1}^{q_{2,n}} \sum_{l=1}^{q_{3,n}} \pi(|b_{jkl}| > b_n)$$

Since $b_{jkl} \sim \mathcal{N}(0, \sigma_{jkl}^2)$, $\frac{1}{\sigma_{jkl}^2} > \frac{1}{\lambda_n}$. By Mill's ratio $\pi(|\frac{b_{jkl}}{\sqrt{\lambda_n}}| > \frac{b_n}{\sqrt{\lambda_n}}) < 2 \frac{\exp\{-b_n^2/2\tilde{\lambda}_n\}}{\sqrt{2\pi b_n^2/\tilde{\lambda}_n}}$, the above quantity is bounded above by $2q_n \frac{\exp\{-b_n^2/2\tilde{\lambda}_n\}}{\sqrt{2\pi b_n^2/\tilde{\lambda}_n}} = 2q_n \frac{\exp\{-4n\varepsilon_n^2\}}{4\sqrt{\pi n\varepsilon_n^2}} \leq \exp\{-2n\varepsilon_n^2\}$ for sufficiently large n , since $\log(q_n)/(n\varepsilon_n^2) \rightarrow 0$ from the assumptions *i)* of Theorem 3 and $n\varepsilon_n^2 \rightarrow \infty$. Condition *b* follows.

Condition *c*. We verify condition *c* for $t = 1$. From Proposition 3, we have,

$$P(|\langle \text{GTRP}(\mathcal{X}), \mathcal{B} \rangle - \langle \mathcal{X}, \mathcal{B}_0 \rangle| < \Delta) > P(X - Y \geq 2).$$

Since $X \sim \text{Poi}(\frac{\Delta_1}{2})$, $Y \sim \text{Poi}(\frac{\lambda}{2})$, $X - Y$ follows Skellam distribution with PMF

$$P(X - Y = k) = \exp\{-(\lambda + \Delta_1)\} \left(\frac{\Delta_1}{\lambda}\right) I_k\left(2\sqrt{\lambda\Delta_1}\right)$$

Plug in λ , Δ_1 and $k = 2$ we have,

$$P(X - Y = 2) = \exp\left\{-\frac{\Delta^2 + \langle \mathcal{X}, \mathcal{B}_0 \rangle^2}{\text{Var}(\langle \text{GTRP}(\mathcal{X}), \mathcal{B} \rangle)}\right\} \left(\frac{\Delta^2}{\langle \mathcal{X}, \mathcal{B}_0 \rangle^2}\right) I_2\left(2\frac{\Delta|\langle \mathcal{X}, \mathcal{B}_0 \rangle|}{\text{Var}(\langle \text{GTRP}(\mathcal{X}), \mathcal{B} \rangle)}\right) \quad (\text{A.49})$$

using the fact that for $z > 0$, $I_k(z) > 2^k z^k \Gamma(k+1)$ (Joshi and Bissu, 1991), we have

$$\begin{aligned} P(X - Y \geq 2) &> P(X - Y = 2) \\ &> \exp\left\{-\frac{\Delta^2 + \langle \mathcal{X}, \mathcal{B}_0 \rangle^2}{\text{Var}(\langle \text{GTRP}(\mathcal{X}), \mathcal{B} \rangle)}\right\} \left(\frac{\Delta}{\langle \mathcal{X}, \mathcal{B}_0 \rangle}\right)^2 2^2 \left(2\frac{\Delta \langle \mathcal{X}, \mathcal{B}_0 \rangle}{\text{Var}(\langle \text{GTRP}(\mathcal{X}), \mathcal{B} \rangle)}\right)^2 \Gamma(3) \\ &> \exp\left\{-\frac{\Delta^2 + \langle \mathcal{X}, \mathcal{B}_0 \rangle^2}{\text{Var}(\langle \text{GTRP}(\mathcal{X}), \mathcal{B} \rangle)}\right\} \frac{2^5 \Delta^4}{\text{Var}(\langle \text{GTRP}(\mathcal{X}), \mathcal{B} \rangle)^2} \\ &> \exp\left\{-\frac{\Delta^2 + \langle \mathcal{X}, \mathcal{B}_0 \rangle^2}{\lambda \|\text{GTRP}(\mathcal{X})\|_F^2}\right\} \frac{2^5 \Delta^4}{\tilde{\lambda}^2 \|\text{GTRP}(\mathcal{X})\|_F^4} > \exp\left\{-\frac{n\varepsilon_n^2}{4}\right\} \end{aligned}$$

where the last inequality follows from

$$\begin{aligned} \exp\left\{-\frac{\Delta^2 + \langle \mathcal{X}, \mathcal{B}_0 \rangle^2}{\lambda \|\text{GTRP}(\mathcal{X})\|_F^2}\right\} &> \exp\left\{-\frac{n\varepsilon_n^2}{8} \frac{8}{n\varepsilon_n^2 \lambda \|\text{GTRP}(\mathcal{X})\|_F^2} \frac{\Delta^2 + K^2}{\lambda \|\text{GTRP}(\mathcal{X})\|_F^2}\right\} \\ &> \exp\left\{-\frac{n\varepsilon_n^2}{8} \frac{8}{n\varepsilon_n^2 B_1 \|\text{GTRP}(\mathcal{X})\|_F^2} \frac{\log(q_n)(1 + K^2)}{\lambda \|\text{GTRP}(\mathcal{X})\|_F^2}\right\} > \exp\left\{-\frac{n\varepsilon_n^2}{8}\right\} \end{aligned}$$

choosing $\Delta = \varepsilon_n^2/(4\eta)$ and assuming $\lambda > B_1/\log(q_n)$ as in *ii)* and $\|\text{GTRP}(\mathcal{X})\|_F^2 > 8(1 + K^2)\log(q_n)/(n\varepsilon_n^2 B_1)$ as in *iii)*, and from

$$\begin{aligned} \frac{2^5 \Delta^4}{\tilde{\lambda}^2 \|\text{GTRP}(\mathcal{X})\|_F^4} &= \exp\left\{-\frac{n\varepsilon_n^2}{8} \left(\frac{8\log(\tilde{\lambda}^2) - 8\log(2^5 \Delta^4)}{n\varepsilon_n^2} + \frac{8\log(\|\text{GTRP}(\mathcal{X})\|_F^4)}{n\varepsilon_n^2}\right)\right\} \\ \exp\left\{-\frac{n\varepsilon_n^2}{8} \left(8\frac{2\log(B) + 2v\log(q_n) - \log(2^5 \Delta^4)}{n\varepsilon_n^2} + \frac{8\log(\|\text{GTRP}(\mathcal{X})\|_F^4)}{n\varepsilon_n^2}\right)\right\} &> \exp\left\{-\frac{n\varepsilon_n^2}{8}\right\} \end{aligned}$$

due to assumptions $\log(q_n)/(n\varepsilon_n^2) \rightarrow 0$ in *i)*, $\tilde{\lambda}_n \leq Bq_n^v$ in *ii)* and $\log(\|\text{GTRP}(\mathcal{X})\|)/(n\varepsilon_n^2) \rightarrow 0$ in *iii)* as $n \rightarrow \infty$. We conclude that for all large n

$$P\left(|\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \frac{\varepsilon_n^2}{4\eta}\right) > \exp\left\{-\frac{n\varepsilon_n^2}{4}\right\}.$$

For $\mathcal{X} = \mathcal{X}_1, \dots, \mathcal{X}_n$, let $\mathcal{S} = \left\{ \mathcal{B} : |\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \frac{\varepsilon_n^2}{4\eta} \right\}$. For $t = 1$,

$$\begin{aligned} d_{t=1} &= \iint f_0 \left(\frac{f_0}{f} - 1 \right) \nu_y(dy) \nu_{\mathcal{X}}(d\mathcal{X}) \\ &= \int E_{y|\mathcal{X}} \left[\frac{f_0}{f}(Y) - 1 \right] \nu_{\mathcal{X}}(d\mathcal{X}) = E_{\mathcal{X}} [g(u^*) (\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \mathcal{X}_i, \mathcal{B}_0 \rangle)] \end{aligned}$$

where the last steps are achieved by first integrating out y and applying mean value theorem. g is a continuous derivative function and u^* is an intermediate point between $\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle$ and $\langle \mathcal{X}_i, \mathcal{B}_0 \rangle$. Since $|\langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \sum_n |b_{jkl,0}| < K$, we can bound u^* by the following,

$$|u^*| < |\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \mathcal{X}_i, \mathcal{B}_0 \rangle| + |\langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \frac{\varepsilon_n^2}{4\eta} + K$$

Choosing η such that $|g(u^*)| < \eta$ in the interval $[-(K+1), (K+1)]$ for all large n , this implies $d_t(f, f_0) < \frac{\varepsilon_n^2}{4}$ is a subset of $\mathcal{S}$, hence confirming condition c .

A.5 Proof of Theorem 4

We show that the three conditions are also satisfied with PARAFAC priors. Following the prior imposed on the margins from the PARAFAC decomposition from Eq.10, we have $\gamma_m^{(d)} \sim \mathcal{N}_{p_m}(\mathbf{0}, \tau \zeta^{(d)} W_m^{(d)})$.

Condition a is easily verified with the same spirits as in the proof of Thm 3.

Condition b.

By PARAFAC decomposition, we have:

$$\pi(|b_{jkl}| \leq b_n) = \pi \left(\left| \sum_{d=1}^D \gamma_{1,j}^{(d)} \gamma_{2,k}^{(d)} \gamma_{3,l}^{(d)} \right| \leq b_n \right) \quad (\text{A.50})$$

$$\geq \pi \left(\sum_{d=1}^D |\gamma_{1,j}^{(d)} \gamma_{2,k}^{(d)} \gamma_{3,l}^{(d)}| \leq b_n \right) \quad (\text{A.51})$$

$$\geq \pi \left(|\gamma_{1,j}^{(d)} \gamma_{2,k}^{(d)} \gamma_{3,l}^{(d)}| \leq \frac{b_n}{D} \right) \quad (\text{A.52})$$

$$\geq \pi \left(|\gamma_{m,j_m}^{(d)}| \leq \left(\frac{b_n}{D} \right)^{1/M} \right) \quad (\text{A.53})$$

Therefore, $\pi(|b_{jkl}| > b_n) \leq \pi \left(|\gamma_{m,j_m}^{(d)}| > \left(\frac{b_n}{D} \right)^{1/M} \right)$. By Mill's ratio $\pi \left(\left| \frac{\gamma_{m,j_m}^{(d)}}{\sqrt{\tilde{\lambda}_n}} \right| > \frac{(\frac{b_n}{D})^{1/M}}{\sqrt{\tilde{\lambda}_n}} \right) < 2 \frac{\exp\{-\frac{(\frac{b_n}{D})^{2/M}}{2\tilde{\lambda}_n}\}}{\sqrt{2\pi(\frac{b_n}{D})^{2/M}/\tilde{\lambda}_n}}$. Let $b_n = D(8\tilde{\lambda}_n n \varepsilon_n^2)^{M/2}$, the results follow from the same arguments used in proof of Thm 3 condition b .

Condition c.

We are interested in a lower bound for

$$P(|\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \Delta_n). \quad (\text{A.54})$$

Notice that

$$\begin{aligned} &P(|\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \Delta_n) \\ &\geq P(|\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle| + |\langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \Delta_n) \\ &\geq P(\|\text{GTRP}(\mathcal{X}_i)\| \|\mathcal{B}\| + K < \Delta_n) \end{aligned}$$

where $|\langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < K$ the first inequality follows from a probabilistic triangular inequality¹ and the second inequality follows Cauchy-Schwartz inequality $|\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle| \leq \|\text{GTRP}(\mathcal{X}_i)\| \|\mathcal{B}\|$. Let $\omega_n = \frac{|\Delta_n - K|}{\|\text{GTRP}(\mathcal{X}_i)\|}$, and the fact that $\|\mathcal{B}\| = \|\sum_{d=1}^D (\gamma_1^{(d)} \circ \dots \circ \gamma_M^d)\| \leq \sum_{d=1}^D \|\gamma_1^{(d)} \circ \dots \circ \gamma_M^d\|$ by triangular inequality, thus

$$\begin{aligned} & P\left(\|\mathcal{B}\| < \frac{|\Delta_n - K|}{\|\text{GTRP}(\mathcal{X}_i)\|}\right) \\ &= P\left(\left\|\sum_{d=1}^D \gamma_1^{(d)} \circ \dots \circ \gamma_M^d\right\| \leq \omega_n\right) \end{aligned} \quad (\text{A.55})$$

$$\geq P\left(\sum_{d=1}^D \|\gamma_1^{(d)} \circ \dots \circ \gamma_M^d\| \leq \omega_n\right) \quad (\text{A.56})$$

$$\geq P\left(\bigcap_{d=1}^D \left\{\|\gamma_1^{(d)} \circ \dots \circ \gamma_M^d\| \leq \frac{\omega_n}{D}\right\}\right) \quad (\text{A.57})$$

$$= \prod_{d=1}^D P\left(\|\gamma_1^{(d)} \circ \dots \circ \gamma_M^d\| \leq \frac{\omega_n}{D}\right) \quad (\text{A.58})$$

$$= \prod_{d=1}^D P\left(\|\gamma_1^{(d)}\| \dots \|\gamma_M^d\| \leq \frac{\omega_n}{D}\right) \quad (\text{A.59})$$

$$\geq \prod_{d=1}^D P\left(\bigcap_{m=1}^M \left\{\|\gamma_m^{(d)}\| \leq \left(\frac{\omega_n}{D}\right)^{1/M}\right\}\right) \quad (\text{A.60})$$

$$= \prod_{d=1}^D \prod_{m=1}^M P\left(\|\gamma_m^{(d)}\| \leq \left(\frac{\omega_n}{D}\right)^{1/M}\right), \quad (\text{A.61})$$

where the inequality from (A.55) to (A.56) follows triangular inequality. From (A.56) to (A.57) is due to the fact that $\bigcap_{d=1}^D \left\{\|\gamma_1^{(d)} \circ \dots \circ \gamma_M^d\| \leq \frac{\omega_n}{D}\right\} \subset \left\{\sum_{d=1}^D \|\gamma_1^{(d)} \circ \dots \circ \gamma_M^d\| \leq \omega_n\right\}$.

From (A.59) to (A.60) is due to the fact that $\bigcap_{m=1}^M \left\{\|\gamma_m^{(d)}\| \leq \left(\frac{\omega_n}{D}\right)^{1/M}\right\} \subset \left\{\|\gamma_1^{(d)} \circ \dots \circ \gamma_M^d\| \leq \frac{\omega_n}{D}\right\}$

Let $\kappa_n = \left(\frac{\omega_n}{D}\right)^{1/M} = \left(\frac{|\Delta_n - K|}{D\|\text{GTRP}(\mathcal{X}_i)\|}\right)^{1/M}$, we need to bound $P\left(\|\gamma_m^{(d)}\| \leq \kappa_n\right)$.

$$\begin{aligned} & P\left(\|\gamma_m^{(d)}\| \leq \kappa_n | \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right) \\ & \geq \prod_{j_m=1}^{q_{m,n}} P\left(|\gamma_{m,j_m}^{(d)}| \leq \frac{\kappa_n}{\sqrt{q_{m,n}}} | \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right) \\ & \geq \prod_{j_m=1}^{q_{m,n}} \left(\frac{2\kappa_n}{\sqrt{q_{m,n} \tau \zeta^{(d)} w_{m,j_m}^{(d)}}} \exp\left\{-\frac{\kappa_n^2}{q_{m,n} \tau \zeta^{(d)} w_{m,j_m}^{(d)}}\right\} \right) \end{aligned}$$

where the last step follows from the fact that $\int_a^b e^{-x^2/2} dx \geq e^{-(a^2+b^2)/2}(b-a)$. Let $\varphi(\kappa_n) = \prod_{j_m=1}^{q_{m,n}} \left(\frac{2\kappa_n}{\sqrt{q_{m,n} \tau \zeta^{(d)} w_{m,j_m}^{(d)}}} \exp\left\{-\frac{\kappa_n^2}{q_{m,n} \tau \zeta^{(d)} w_{m,j_m}^{(d)}}\right\} \right)$. We want to show $-\log \varphi(\kappa_n) < \frac{n\varepsilon_n^2}{a}$.

Note that

$$P\left(\|\gamma_m^{(d)}\| \leq \kappa_n | \tau, \zeta^{(d)}\right)$$

¹The following result returns the inequality. Let $Q = \{\omega : |A(\omega) - B(\omega)| < \Delta\}$ and $R = \{\omega : |A(\omega)| + |B(\omega)| < \Delta\}$ be two events. Note that $Q = Q \cap (R \cup R^C) = (Q \cap R) \cup (Q \cap R^C)$, and $Q \cap R = \{\omega : |A(\omega) - B(\omega)| < \Delta \text{ and } |A(\omega)| + |B(\omega)| < \Delta\} = \{\omega : |A(\omega)| + |B(\omega)| < \Delta\} = R$ from standard triangular inequality. Thus $R \subset Q$, and $P(R) < P(Q)$.

$$\begin{aligned}
&= \mathbb{E} \left[P \left(\|\gamma_m^{(d)}\| \leq \kappa_n | \tau, \zeta^{(d)}, w_{m,j_m}^{(d)} \right) \right] \\
&\geq \left(\frac{2\kappa_n}{\sqrt{q_{m,n}\tau\zeta^{(d)}}} \right)^{q_{m,n}} \prod_{j_m=1}^{q_{m,n}} \mathbb{E} \left[\left(\frac{1}{\sqrt{w_{m,j_m}^{(d)}}} \exp \left\{ -\frac{\kappa_n^2}{q_{m,n}\tau\zeta^{(d)}w_{m,j_m}^{(d)}} \right\} \right) \right] \\
&= \left(\frac{2\kappa_n\lambda_m^{(d)2}}{2\sqrt{q_{m,n}\tau\zeta^{(d)}}} \right)^{q_{m,n}} \prod_{j_m=1}^{q_{m,n}} \int \left(\frac{1}{\sqrt{w_{m,j_m}^{(d)}}} \exp \left\{ -\frac{\kappa_n^2}{q_{m,n}\tau\zeta^{(d)}w_{m,j_m}^{(d)}} - \frac{\lambda_m^{(d)2}w_{m,j_m}^{(d)}}{2} \right\} \right) dw_{m,j_m}^{(d)} \\
&= \left(\frac{\kappa_n\lambda_m^{(d)2}}{\sqrt{q_{m,n}\tau\zeta^{(d)}}} \right)^{q_{m,n}} \exp \left\{ -\lambda_m^{(d)}\kappa_n\sqrt{\frac{2q_{m,n}}{\tau\zeta^{(d)}}} \right\}
\end{aligned}$$

Following similar reasoning as in Guhaniyogi et al. (2017) we move on to integrate out $\lambda_m^{(d)}$, τ and $\zeta^{(d)}$, and we end up with the following expression

$$\begin{aligned}
&P \left(\|\gamma_m^{(d)}\| \leq \kappa_n, d = 1, \dots, D, m = 1, \dots, M \right) \\
&\geq \frac{\lambda_2^{\lambda_1} \Gamma(Da)}{\Gamma(\lambda_1) \Gamma(a)^D} \prod_{m=1}^M \prod_{d=1}^D \left[\left(\frac{\kappa_n}{\sqrt{q_{m,n}b_{\lambda,d}}} \right)^{q_{m,n}} \frac{\Gamma(a + a_{\lambda,d} \frac{M}{2})}{\Gamma(a_{\lambda,d})} \right] \\
&\quad \prod_{m=1}^M \prod_{d=1}^D \frac{1}{\left(\frac{\sqrt{2q_{m,n}\kappa_n}}{b_{\lambda,d}} + 1 \right)^{q_{m,n} + a_{\lambda,d}}} \frac{\exp\{-\lambda_2\}}{(\lambda_1 + \sum_{d=1}^D a_{\lambda,d} \frac{M}{2})} \frac{\prod_{d=1}^D [\Gamma(a + a_{\lambda,d} \frac{M}{2})]}{\Gamma(Da + \frac{M}{2} \sum_{d=1}^D a_{\lambda,d})}
\end{aligned}$$

Let $C_1 = \frac{\lambda_2^{\lambda_1} \Gamma(Da)}{\Gamma(\lambda_1) \Gamma(a)^D} \frac{\exp\{-\lambda_2\}}{(\lambda_1 + \sum_{d=1}^D a_{\lambda,d} \frac{M}{2})} \frac{\prod_{d=1}^D [\Gamma(a + a_{\lambda,d} \frac{M}{2})]}{\Gamma(Da + \frac{M}{2} \sum_{d=1}^D a_{\lambda,d})}$, then we have

$$\begin{aligned}
&-\log P \left(\|\gamma_m^{(d)}\| \leq \kappa_n \right) \leq -\log C_1 \\
&+ \sum_{m=1}^M \sum_{d=1}^D \left(q_{m,n} \left[-\log \kappa_n + \frac{1}{2} \log q_{m,n} + \log b_{\lambda,d} \right] - \log \Gamma(q_{m,n} + a_{\lambda,d}) + \log \Gamma(a_{\lambda,d}) \right) \\
&+ \sum_{m=1}^M \sum_{d=1}^D (q_{m,n} + a_{\lambda,d}) \log \left(\frac{2\sqrt{q_{m,n}\kappa_n}}{b_{\lambda,d}} + 1 \right) \\
&= \frac{n\varepsilon_n^2}{4} \left(-\frac{4 \log C_1}{n\varepsilon_n^2} \right. \\
&+ \sum_{m=1}^M \sum_{d=1}^D \left(4 \left[-\frac{q_{m,n} \log \kappa_n}{n\varepsilon_n^2} + \frac{q_{m,n} \log q_{m,n}}{2n\varepsilon_n^2} + \frac{q_{m,n} \log b_{\lambda,d}}{n\varepsilon_n^2} \right] - \frac{4 \log \Gamma(q_{m,n} + a_{\lambda,d})}{n\varepsilon_n^2} + \frac{4 \log \Gamma(a_{\lambda,d})}{n\varepsilon_n^2} \right) \\
&+ \sum_{m=1}^M \sum_{d=1}^D \frac{4(q_{m,n} + a_{\lambda,d})}{n\varepsilon_n^2} \log \left(\frac{2\sqrt{q_{m,n}\kappa_n}}{b_{\lambda,d}} + 1 \right) \Big)
\end{aligned}$$

Notice that $-\frac{\log C_1}{n\varepsilon_n^2} \rightarrow 0$ as $n\varepsilon_n^2 \rightarrow \infty$. By plug in $\kappa_n = (\frac{|\Delta_n - K|}{D\|\text{GTRP}(\mathcal{X}_i)\|})^{1/M}$, we have that

$$\begin{aligned}
&-\sum_{m=1}^M \sum_{d=1}^D \frac{q_{m,n} \log \kappa_n}{n\varepsilon_n^2} \\
&= -\frac{D [\log |\Delta_n - K| - \log (\|\text{GTRP}(\mathcal{X}_i)\|) - \log D]}{M} \frac{\sum_{m=1}^M q_{m,n}}{n\varepsilon_n^2} \\
&= -\frac{D \log (K - \varepsilon_n^2)}{M} \frac{\sum_{m=1}^M q_{m,n}}{n\varepsilon_n^2} + \frac{D \log (\|\text{GTRP}(\mathcal{X}_i)\|) + D \log D}{M} \frac{\sum_{m=1}^M q_{m,n}}{n\varepsilon_n^2} \\
&> -\frac{D \log K}{M} \frac{\sum_{m=1}^M q_{m,n}}{n\varepsilon_n^2} + C
\end{aligned}$$

choosing $\Delta_n = \varepsilon_n^2$, and by assumption (iv). From assumption (v) it follows that

$$\frac{\log K \sum_{m=1}^M q_{m,n}}{n\varepsilon_n^2} \rightarrow 0.$$

and $\sum_{m=1}^M q_{m,n} \log q_{m,n} / n\varepsilon_n^2 \rightarrow 0$, which implies $\sum_{m=1}^M q_{m,n} / n\varepsilon_n^2 \rightarrow 0$, $\sum_{m=1}^M \log q_{m,n} / n\varepsilon_n^2 \rightarrow 0$ and $\frac{4(q_{m,n} + a_{\lambda,d})}{n\varepsilon_n^2} \log \left(\frac{2\sqrt{q_{m,n}\kappa_n}}{b_{\lambda,d}} + 1 \right) \rightarrow 0$. By the Stirling approximation of the Gamma function and from assumption (ii), it follows $\frac{4 \log \Gamma(q_{m,n} + a_{\lambda,d})}{n\varepsilon_n^2} \rightarrow 0$. Thus we can claim that $-\log P \left(\|\gamma_m^{(d)}\| \leq \kappa_n \right) \leq \frac{n\varepsilon_n^2}{4}$, thus $P \left(\|\gamma_m^{(d)}\| \leq \kappa_n \right) \geq \exp \left\{ -\frac{n\varepsilon_n^2}{4} \right\}$, which implies

$$P \left(|\langle \text{GTRP}(\mathcal{X}_i), \mathcal{B} \rangle - \langle \mathcal{X}_i, \mathcal{B}_0 \rangle| < \Delta \right) \geq \exp \left\{ -\frac{n\varepsilon_n^2}{4} \right\}.$$

The result follows from the same arguments used in the proof of Proposition 2.

B Full conditional distributions

B.1 PARAFAC priors

Given the PARAFAC priors, the posterior of the unknowns of the model is given by

$$p(\gamma_m^{(d)}, \sigma^2, \mu, w_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)} \mid \mathbf{y}, \mathcal{X}) \quad (\text{B.1})$$

We adopt the MCMC procedure based on the Gibbs sampling algorithm to sample the unknowns from 3 blocks to reduce autocorrelation.

B.1.1 Block 1: Sampling $\zeta^{(d)}$ and τ from $p(\zeta^{(d)}, \tau \mid \gamma, \mathbf{w})$

$$\begin{aligned} p(\zeta^{(d)} \mid \gamma, \tau, \mathbf{w}) &\propto p(\gamma \mid \zeta, \tau, \mathbf{w}) p(\zeta) \\ &\propto \prod_{d=1}^D \prod_{m=1}^M \zeta^{(d)-\frac{p_m}{2}} \exp \left\{ -\frac{1}{2} \gamma_m^{(d)T} \frac{W_m^{(d)-1}}{\tau \zeta^{(d)}} \gamma_m^{(d)} \right\} \prod_{d=1}^D \zeta^{(d)\alpha-1} \\ &= \prod_{d=1}^D \zeta^{(d)-\sum_{m=1}^M p_m/2+\alpha-1} \exp \left\{ -\frac{1}{2\tau \zeta^{(d)}} \sum_{m=1}^M \gamma_m^{(d)T} W_m^{(d)-1} \gamma_m^{(d)} \right\} \\ &\sim \mathcal{GiG} \left(\alpha - \frac{\sum_{m=1}^M p_m}{2}, 0, \frac{\sum_{m=1}^M \gamma_m^{(d)T} W_m^{(d)-1} \gamma_m^{(d)}}{\tau} \right) \\ &\sim \mathcal{IG} \left(\frac{\sum_{m=1}^M p_m}{2} - \alpha, \frac{\sum_{m=1}^M \gamma_m^{(d)T} W_m^{(d)-1} \gamma_m^{(d)}}{2\tau} \right) \end{aligned}$$

$$\begin{aligned} p(\tau \mid \gamma, \zeta, \mathbf{w}) &\propto p(\gamma \mid \zeta, \tau, \mathbf{w}) p(\tau) \\ &\propto \prod_{d=1}^D \prod_{m=1}^M \tau^{-\frac{p_m}{2}} \exp \left\{ -\frac{1}{2} \gamma_m^{(d)T} \frac{W_m^{(d)-1}}{\tau \zeta^{(d)}} \gamma_m^{(d)} \right\} \tau^{a_\tau-1} \exp \{-b_\tau \tau\} \\ &= \tau^{a_\tau - \frac{D \sum_{m=1}^M p_m}{2} - 1} \exp \left\{ -\frac{1}{2\tau} \sum_{d=1}^D \frac{\sum_{m=1}^M \gamma_m^{(d)T} W_m^{(d)-1} \gamma_m^{(d)}}{\zeta^{(d)}} - b_\tau \tau \right\} \\ &\sim \mathcal{GiG} \left(a_\tau - \frac{D \sum_{m=1}^M p_m}{2}, 2b_\tau, \sum_{d=1}^D \frac{\sum_{m=1}^M \gamma_m^{(d)T} W_m^{(d)-1} \gamma_m^{(d)}}{\zeta^{(d)}} \right) \end{aligned}$$

B.1.2 Block 2: Sampling $\lambda_m^{(d)}$ and $w_{m,j_m}^{(d)}$ from $p(\lambda_m^{(d)}, w_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)}, \tau, \zeta^{(d)})$

Notice that by the construction of the prior distributions, $\gamma_{m,j_m}^{(d)}$ follows a double exponential distribution given $\lambda_m^{(d)}$, τ , $\zeta^{(d)}$, that is $\gamma_{m,j_m}^{(d)} \sim \mathcal{DE}\left(0, \sqrt{\tau\zeta^{(d)}}/\lambda_m^{(d)}\right)$. The full conditional of $\lambda_m^{(d)}$ can be written as

$$\begin{aligned} p\left(\lambda_m^{(d)} | \gamma_{m,j_m}^{(d)}, \tau, \zeta^{(d)}\right) &\propto \pi(\lambda_m^{(d)}) p\left(\gamma_{m,j_m}^{(d)} | \lambda_m^{(d)}, \tau, \zeta^{(d)}\right) \\ &\propto \left(\tau\zeta^{(d)}\right)^{-\frac{p_m}{2}} \left(\lambda_m^{(d)}\right)^{a_\lambda + p_m - 1} \exp\left\{-\left(\frac{\sum_{j_m=1}^{p_m} |\gamma_{m,j_m}^{(d)}|}{\sqrt{\tau\zeta^{(d)}}} + b_\lambda\right) \lambda_m^{(d)}\right\} \\ &\propto \mathcal{Ga}\left(a_\lambda + p_m, \sum_{j_m=1}^{p_m} |\gamma_{m,j_m}^{(d)}| / \sqrt{\tau\zeta^{(d)}} + b_\lambda\right) \end{aligned}$$

The full conditional for $w_{m,j_m}^{(d)}$ is

$$\begin{aligned} p\left(w_{m,j_m}^{(d)} | \gamma_{m,j_m}^{(d)}, \lambda_m^{(d)}, \tau, \zeta^{(d)}\right) &\propto \pi\left(w_{m,j_m}^{(d)}\right) p\left(\gamma_{m,j_m}^{(d)} | \lambda_m^{(d)}, \tau, \zeta^{(d)}, w_{m,j_m}^{(d)}\right) \\ &\propto w_{m,j_m}^{(d)\frac{1}{2}-1} \exp\left\{-\frac{1}{2}\left(\lambda_m^{(d)2} w_{m,j_m}^{(d)} + \frac{\gamma_{m,j_m}^{(d)2}}{\tau\zeta^{(d)} w_{m,j_m}^{(d)}}\right)\right\} \\ &\propto \mathcal{GiG}\left(1/2, \lambda_m^{(d)2}, \gamma_{m,j_m}^{(d)2}/\tau\zeta^{(d)}\right) \end{aligned}$$

B.1.3 Block 3: Sampling $\gamma_m^{(d)}, \mu, \sigma^2$

$$\begin{aligned} p(\gamma_m^{(d)} | \mathbf{y}, \mathcal{X}, \tau, \zeta, \mathbf{w}, \mu, \sigma^2) &\propto p(\mathbf{y} | \gamma_m^{(d)}, \mathcal{X}, \tau, \zeta, \mathbf{w}, \mu, \sigma^2) p(\gamma_m^{(d)}) \\ &\propto \prod_{t=1}^T \exp\left\{-\frac{1}{2} \frac{(y_t - \mu - \langle \mathcal{B}, \mathcal{X}_t \rangle)^2}{\sigma^2}\right\} \exp\left\{-\frac{1}{2\tau\zeta^{(d)}} \gamma_m^{(d)T} W_m^{(d)-1} \gamma_m^{(d)}\right\} \end{aligned}$$

notice that

$$\begin{aligned} \langle \mathcal{B}, \mathcal{X}_t \rangle &= \langle \mathcal{B}^{(d)}, \mathcal{X}_t \rangle + \sum_{d' \neq d}^D \langle \mathcal{B}^{(d')}, \mathcal{X}_t \rangle \\ &= \gamma_m^{(d)T} \left(\mathcal{X}_t \times_1 \gamma_1^{(d)} \cdots \times_{m-1} \gamma_{m-1}^{(d)} \times_{m+1} \gamma_{m+1}^{(d)} \cdots \times_M \gamma_M^{(d)} \right) + \sum_{d' \neq d}^D \langle \mathcal{B}^{(d')}, \mathcal{X}_t \rangle \\ &= \gamma_m^{(d)T} \psi_{mt}^{(d)} + R_t^{(d)} \end{aligned}$$

where

$$\begin{aligned} \psi_{mt}^{(d)} &= \mathcal{X}_t \times_1 \gamma_1^{(d)} \cdots \times_{m-1} \gamma_{m-1}^{(d)} \times_{m+1} \gamma_{m+1}^{(d)} \cdots \times_M \gamma_M^{(d)} \\ R_t^{(d)} &= \sum_{d' \neq d}^D \langle \mathcal{B}^{(d')}, \mathcal{X}_t \rangle. \end{aligned}$$

The quadratic term in the likelihood becomes

$$\begin{aligned} (y_t - \mu - \langle \mathcal{B}, \mathcal{X}_t \rangle)^2 &= \left(y_t - \mu - R_t^{(d)}\right)^2 - 2 \left(y_t - \mu - R_t^{(d)}\right) \gamma_m^{(d)T} \psi_{mt}^{(d)} + \gamma_m^{(d)T} \psi_{mt}^{(d)} \psi_{mt}^{(d)T} \gamma_m^{(d)} \end{aligned}$$

$$= (\tilde{y}_t^{(d)})^2 - 2\tilde{y}_t^{(d)}\gamma_m^{(d)T}\psi_{mt}^{(d)} + \gamma_m^{(d)T}\psi_{mt}^{(d)}\psi_{mt}^{(d)T}\gamma_m^{(d)}$$

where $\tilde{y}_t^{(d)} = y_t - \mu - R_t^{(d)}$.

Then we have the full conditional for $\gamma_m^{(d)}$

$$\begin{aligned} p(\gamma_m^{(d)} \mid \mathbf{y}, \mathcal{X}, \tau, \zeta, \mathbf{w}, \mu, \sigma^2) \\ \propto \exp \left\{ -\frac{1}{2\sigma^2} \left[\gamma_m^{(d)T} \sum_{t=1}^T \psi_{mt}^{(d)} \psi_{mt}^{(d)T} \gamma_m^{(d)} - 2\gamma_m^{(d)T} \sum_{t=1}^T \tilde{y}_t^{(d)} \psi_{mt}^{(d)} \right] - \frac{1}{2\tau\zeta^{(d)}} \gamma_m^{(d)T} W_m^{(d)-1} \gamma_m^{(d)} \right\} \\ \propto \exp \left\{ -\frac{1}{2} \left[\gamma_m^{(d)T} \left(\frac{\sum_{t=1}^T \psi_{mt}^{(d)} \psi_{mt}^{(d)T}}{\sigma^2} + \frac{W_m^{(d)-1}}{\tau\zeta^{(d)}} \right) \gamma_m^{(d)} - 2\gamma_m^{(d)T} \frac{\sum_{t=1}^T \tilde{y}_t^{(d)} \psi_{mt}^{(d)}}{\sigma^2} \right] \right\} \\ \sim \mathcal{MN}_{p_m}(\boldsymbol{\mu}^*, \Sigma^*) \end{aligned}$$

where

$$\begin{aligned} \Sigma^* &= \left(\frac{\sum_{t=1}^T \psi_{mt}^{(d)} \psi_{mt}^{(d)T}}{\sigma^2} + \frac{W_m^{(d)-1}}{\tau\zeta^{(d)}} \right)^{-1} \\ \boldsymbol{\mu}^* &= \left(\frac{\sum_{t=1}^T \psi_{mt}^{(d)} \psi_{mt}^{(d)T}}{\sigma^2} + \frac{W_m^{(d)-1}}{\tau\zeta^{(d)}} \right)^{-1} \frac{\sum_{t=1}^T \tilde{y}_t^{(d)} \psi_{mt}^{(d)}}{\sigma^2} \end{aligned}$$

The full conditional of σ^2 can be written as:

$$\begin{aligned} p(\sigma^2 \mid \mathbf{y}, \mathcal{X}, \mu, \gamma) &\propto p(\mathbf{y} \mid \mathcal{X}, \mu, \gamma, \sigma^2) p(\sigma^2) \\ &\propto (\sigma^2)^{-(a_\sigma + \frac{T}{2})-1} \exp \left\{ -\frac{1}{\sigma^2} \left(\frac{1}{2} \sum_{t=1}^T (y_t - \langle \mathcal{B}, \mathcal{X}_t \rangle - \mu)^2 + b_\sigma \right) \right\}, \end{aligned}$$

which is the kernel of the IG distribution $\mathcal{IG}(a_\sigma^*, b_\sigma^*)$, where $a_\sigma^* = a_\sigma + \frac{T}{2}$ and $b_\sigma^* = \frac{1}{2} \sum_{t=1}^T (y_t - \langle \mathcal{B}, \mathcal{X}_t \rangle - \mu)^2 + b_\sigma$. Finally, let $\mu^* = \sum_{t=1}^T (y_t - \langle \mathcal{B}, \mathcal{X}_t \rangle) \sigma_\mu^{*2} / \sigma^2$ and $\sigma_\mu^{*2} = (T/\sigma^2 + 1/\sigma_\mu^2)^{-1}$, the full conditional of μ is:

$$\begin{aligned} p(\mu \mid \mathbf{y}, \mathcal{X}, \gamma, \sigma^2) &\propto p(\mathbf{y} \mid \mathcal{X}, \mu, \gamma, \sigma^2) \pi(\mu) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left[T\mu^2 - 2\mu \sum_{t=1}^T (y_t - \langle \mathcal{B}, \mathcal{X}_t \rangle) \right] - \frac{1}{2} \frac{\mu^2}{\sigma_\mu^2} \right\} \\ &= \exp \left\{ -\frac{1}{2} \left[\left(\frac{T}{\sigma^2} + \frac{1}{\sigma_\mu^2} \right) \mu^2 - 2\mu \frac{\sum_{t=1}^T (y_t - \langle \mathcal{B}, \mathcal{X}_t \rangle)}{\sigma^2} \right] \right\} \propto \mathcal{N}(\mu^*, \sigma_\mu^{*2}). \end{aligned}$$

B.2 Gaussian priors

Given the Gaussian prior for the tensor coefficients specified in Theorem 3, we further more assume that $\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma)$ and $\mu \sim \mathcal{N}(0, \sigma_\mu^2)$, w.o.l.g we assume the tensor coefficient is a mode-2 tensor, then we have the following full conditionals for the tensor coefficients, σ^2 and μ :

$$\begin{aligned} p(\mathcal{B}_{\text{vec}} \mid \mathbf{y}, X, \boldsymbol{\mu}, \sigma^2) &\propto p(\mathbf{y} \mid \mathcal{B}_{\text{vec}}, X) p(\mathcal{B}_{\text{vec}}) \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\mu} - X\mathcal{B}_{\text{vec}})^\top (\mathbf{y} - \boldsymbol{\mu} - X\mathcal{B}_{\text{vec}}) \right\} \exp \left\{ -\frac{1}{2} \mathcal{B}_{\text{vec}}^\top (\Sigma_1 \otimes \Sigma_2)^{-1} \mathcal{B}_{\text{vec}} \right\} \\ &\sim \mathcal{MN}(\boldsymbol{\mu}_{\mathcal{B}_{\text{vec}}}, \Sigma_{\mathcal{B}_{\text{vec}}}) \end{aligned}$$

where $\mathcal{B}_{\text{vec}}$ is the vectorized tensor coefficient $\mathcal{B}$ and X is the matrix obtained stacking vertically vectorized covariate tensors $\text{vec}(\mathcal{X}_t)^\top$, $t = 1, \dots, T$, $\boldsymbol{\mu} = \boldsymbol{\mu}_T$, $\Sigma_{\mathcal{B}_{\text{vec}}} = (X^\top X / \sigma^2 + (\Sigma_1 \otimes \Sigma_2)^{-1})^{-1}$ and $\boldsymbol{\mu}_{\mathcal{B}_{\text{vec}}} = \Sigma_{\mathcal{B}_{\text{vec}}} X^\top (\mathbf{y} - \boldsymbol{\mu}) / \sigma^2$.

$$\begin{aligned}
p(\sigma^2 \mid \mathbf{y}, X, \mathcal{B}_{\text{vec}}, \boldsymbol{\mu}) &\propto p(\mathbf{y} \mid X, \boldsymbol{\mu}, \mathcal{B}_{\text{vec}}, \sigma^2) p(\sigma^2) \\
&\propto (\sigma^2)^{\frac{T}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\mu} - X\mathcal{B}_{\text{vec}})^\top (\mathbf{y} - \boldsymbol{\mu} - X\mathcal{B}_{\text{vec}}) \right\} (\sigma^2)^{-a_\sigma - 1} \exp \left\{ -\frac{b_\sigma}{\sigma^2} \right\} \\
&\sim \mathcal{IG}(a_\sigma^*, b_\sigma^*)
\end{aligned}$$

where $a_\sigma^* = a_\sigma + T/2$ and $b_\sigma^* = b_\sigma + (\mathbf{y} - \boldsymbol{\mu})^\top (\mathbf{y} - \boldsymbol{\mu})/2$.

$$\begin{aligned}
p(\boldsymbol{\mu} \mid \mathbf{y}, X, \mathcal{B}_{\text{vec}}, \sigma^2) &\propto p(\mathbf{y} \mid \boldsymbol{\mu}, X, \mathcal{B}_{\text{vec}}, \sigma^2) p(\boldsymbol{\mu}) \\
&\propto \prod_{t=1}^T \exp \left\{ -\frac{1}{2\sigma^2} (y_t - \mu - \mathcal{X}_t^\top \mathcal{B}_{\text{vec}})^2 \right\} \exp \left\{ -\frac{\mu^2}{2\sigma_\mu^2} \right\} \\
&\sim \mathcal{N}(\mu^*, \sigma_\mu^{*2})
\end{aligned}$$

where $\mu^* = (T/\sigma^2 + 1/\sigma_\mu^2)^{-1} \sum_{t=1}^T (y_t - \mathcal{X}_t^\top \mathcal{B}_{\text{vec}})/\sigma^2$ and $\sigma_\mu^{*2} = (T/\sigma^2 + 1/\sigma_\mu^2)^{-1}$.

C Further numerical results

In this section, we provide further illustration of the effectiveness of the Bayesian compressed tensor regression model proposed in Section 2.

C.1 Sample size

We consider three different simulation settings. In each setting, a different 20×20 true tensor coefficient is used to generate the $n = 1,500$ i.i.d. samples. The tensor covariates, which are also 20×20 , are drawn i.i.d. from the standard normal distribution. The simulated results are presented in Fig. C.1.

Parameter estimation is based on the first 1,000 observations, and out-of-sample forecasts are generated for the remaining 500 samples. The following hyper-parameter setting is considered: $D = 5$, $\alpha = D^{-2}$, $a_\tau = 3$, $b_\tau = 100$, $a_\lambda = 20$, $b_\lambda = 2$, $a_\sigma = 3$, $b_\sigma = 1$, $\sigma_\mu^2 = 1$. We ran the Gibbs sampler for 1,000 iterations and removed 200 burn-in samples.

C.2 Non-structured coefficients

To explore the effects of random projection on tensor coefficients without underlying structure, unlike the settings in previous simulations, we carry out further simulations with the true coefficients, where the entries are i.i.d. drawn from $\{0, 1\}$ at sparsity levels of 75%, 50%, and 25%.

Fig. C.3 shows the scatter plots of predicted data against the actual data across different random projection methods for coefficients with different sparsity levels using compression rate = 0.36, training sample size = 1000, and $\psi = 3$. When the sparsity level of the true coefficients is moderate (25% and 50%), mode-wise random projection and mode-wise random projection with mode preservation still outperform the tensor-wise random projection as in the case of coefficients with some underlying structures. However, when the true coefficients become highly sparse (75%), the performance of the different random projections becomes very close, with tensor-wise random projection slightly outperforming mode-wise random projection. This can also be seen in Fig. C.2, which shows the RMSE across different random projection methods for different true coefficients.

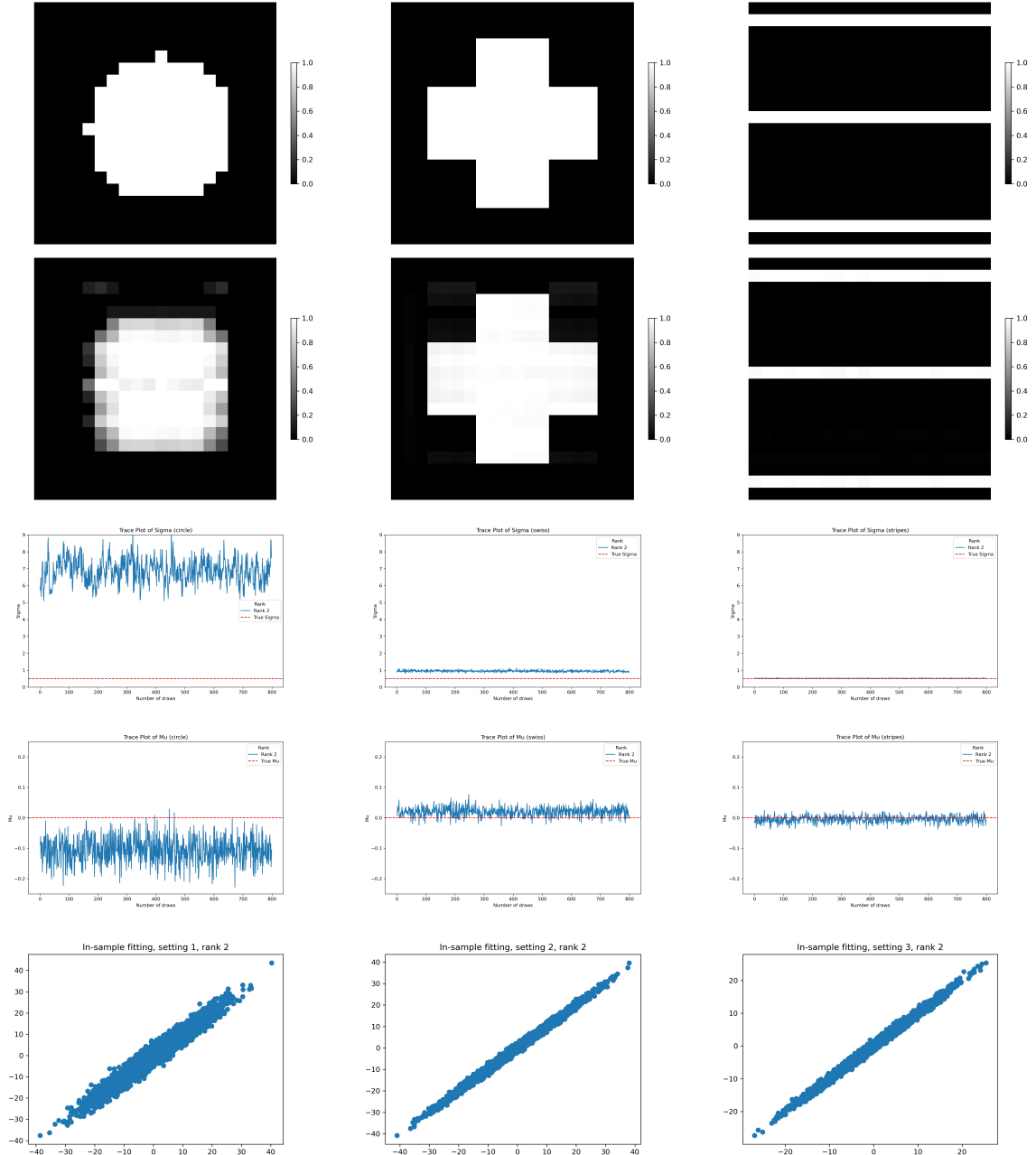


Figure C.1: Simulation results for Bayesian tensor regression. First row: true coefficients. Second row: estimated coefficients. Third and fourth row: trace plots of σ^2 and μ , true values are the red dashed lines. Fifth row: scatter plots for in-sample fitting, true values (horizontal axis) versus fitted values (vertical axis).

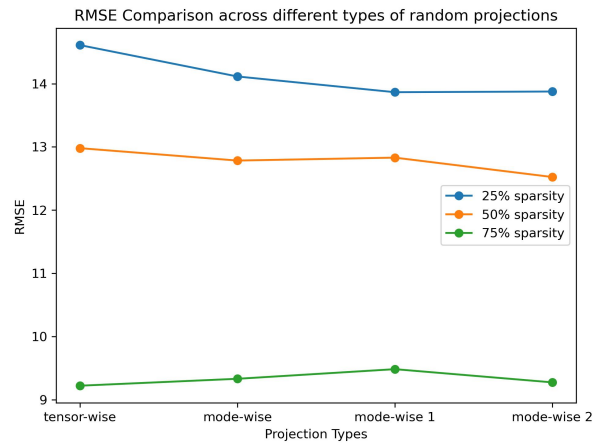


Figure C.2: RMSE (vertical axis) comparison across different random projection methods (horizontal axis) and coefficients with different sparsity levels shown as lines in different colors (25%: blue, 50%: yellow, 75%: green).

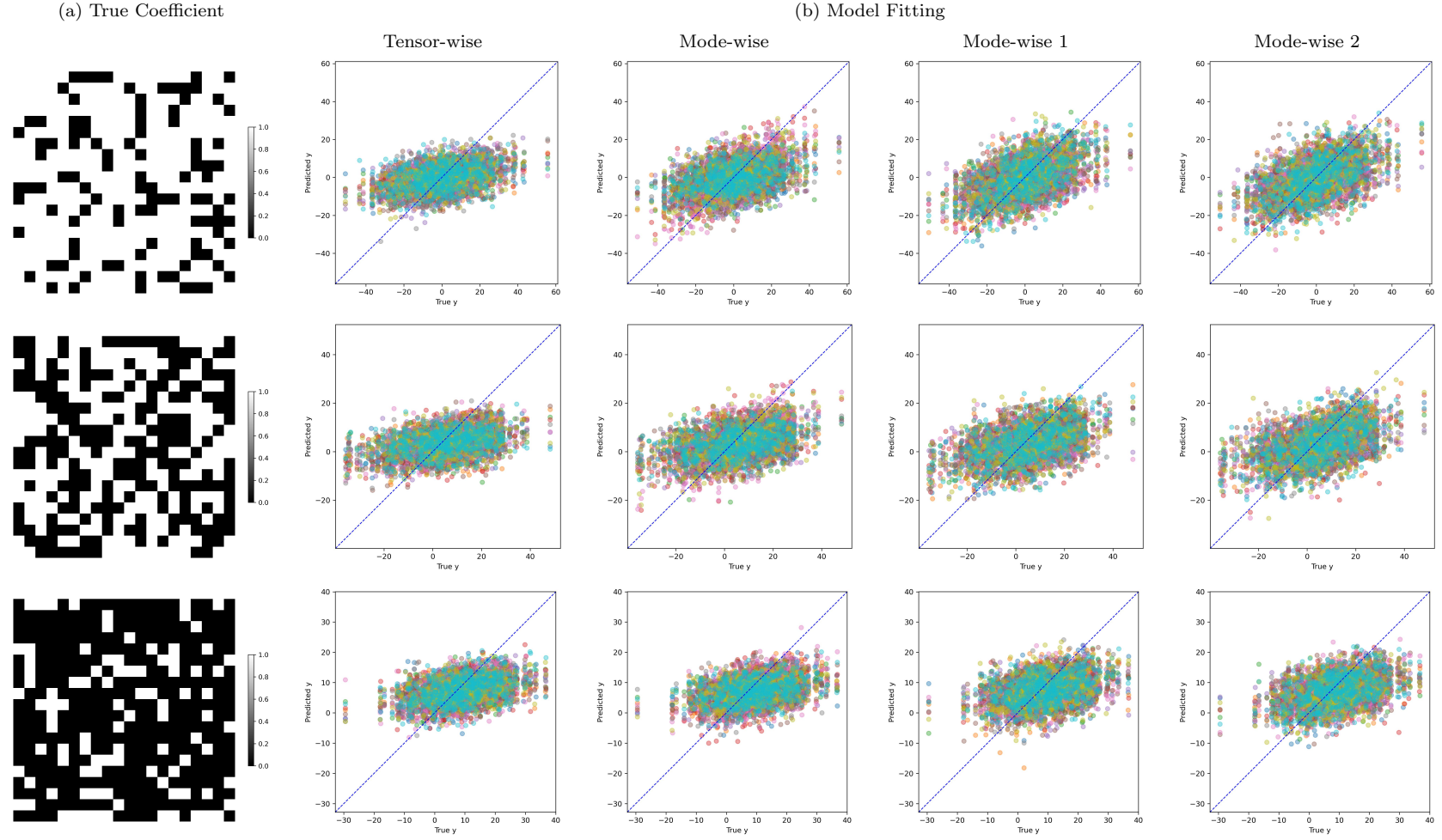


Figure C.3: Scatter plots of actual data versus the predicted for three sets of coefficients with no underlying structures at sparsity levels of 25%, 50%, and 75%. True coefficients are shown in panel (a) and forecasts are shown in panel (b). In each scatter plot: actual data (horizontal axis) against the predicted data (vertical axis) for different levels of sparsity (rows) and different types of random projections (columns), using $L = 10$ independent projection matrices of the same type (colors) for each simulation. In each experiment: training sample size: $n = 1000$, compression rate: 0.36, $\psi = 3$.

D Further empirical results

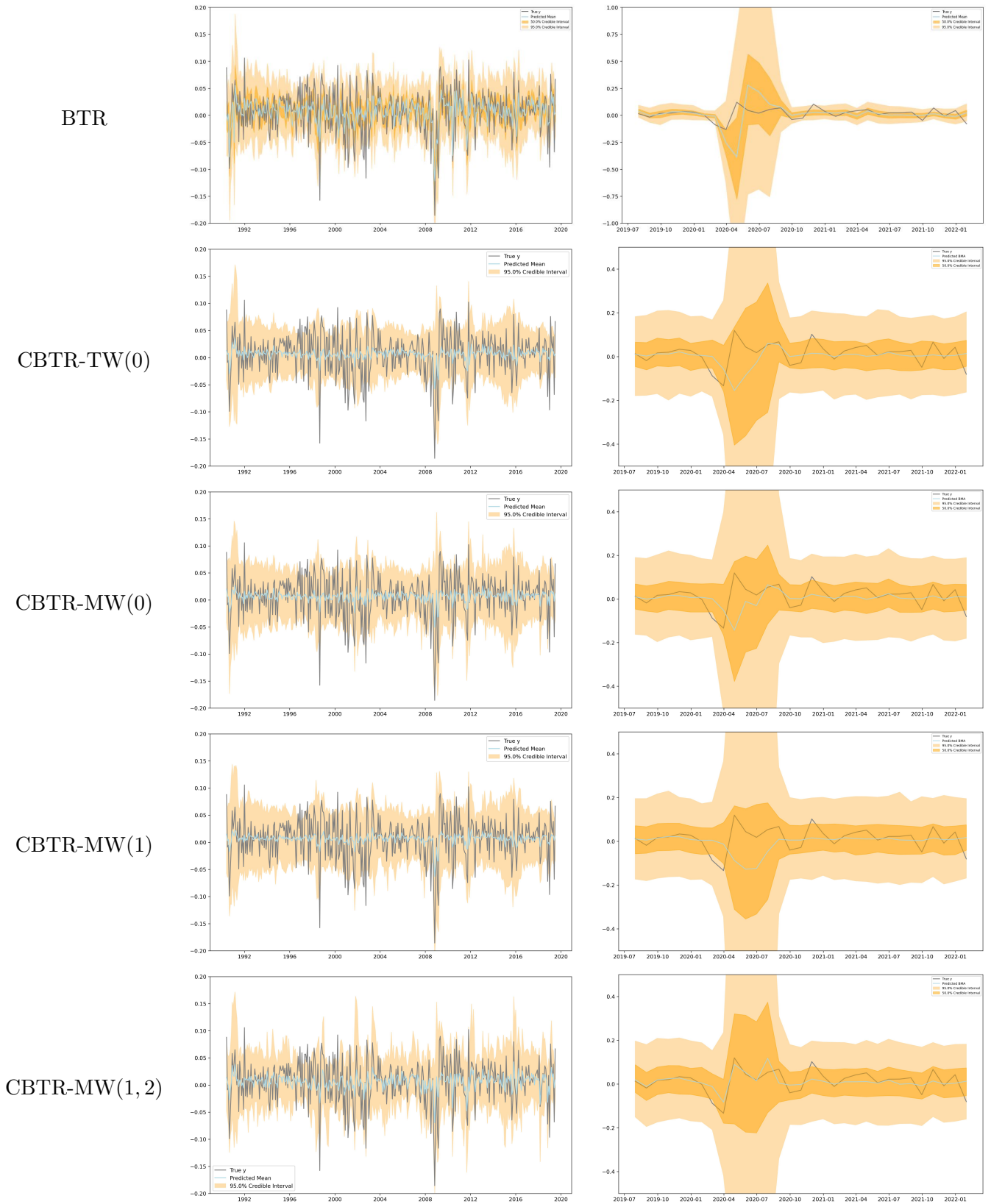


Figure D.1: Fitting comparison between BTR and CBTR with different random projection methods. First column: in-sample fitting. Second column: out-of-sample prediction. Actual data are shown in gray solid line, predicted values are shown in blue solid line, light and dark orange colors represent 95% and 50% credible intervals, respectively.

References

- 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)*.
- Batselier, K. and Wong, N. (2017). A constructive arbitrary-degree Kronecker product decomposition of tensors. *Numerical Linear Algebra with Applications*, 24(5):e2097.
- 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.
- 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.
- Dasgupta, S. (1999). Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE.
- Dasgupta, S. (2013). Experiments with random projection. *arXiv preprint arXiv:1301.3849*.
- 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.
- Farahmand, A.-m., Pourazarm, S., and Nikovski, D. (2017). Random projection filter bank for time series data. *Advances in Neural Information Processing Systems*, 30.
- Feng, L. and Yang, G. (2024). Deep Kronecker network. *Biometrika*, 111(2):707–714.
- Geppert, L. N., Ickstadt, K., Munteanu, A., Quedenfeld, J., and Sohler, C. (2017). Random projections for Bayesian regression. *Statistics and Computing*, 27(1):79–101.
- Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in markov chain monte carlo. *Technical Report 568*.
- 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.
- 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(79):1–31.
- Hackbusch, W. (2019). *Tensor Spaces and Numerical Tensor Calculus*, volume 56 of *Springer Series in Computational Mathematics*. Springer International Publishing, Cham.
- Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083.
- 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.
- Joshi, C. and Bissu, S. (1991). Some inequalities of Bessel and modified Bessel functions. *Journal of the Australian Mathematical Society*, 50(2):333–342.
- 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., 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. *arXiv preprint arXiv:2103.00025*.
- Li, P. and Li, X. (2023). Differential privacy with random projections and sign random projections. *arXiv preprint arXiv:2306.01751*.
- Luo, Y. and Griffin, J. E. (2025). Bayesian inference of vector autoregressions with tensor decompositions. *Journal of Business & Economic Statistics*, pages 1–29.
- Mathai, A. M., Saxena, R. K., and Haubold, H. J. (2010). *The H-function: theory and applications*. Springer, New York.
- 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.

- 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.
- Rakhshan, B. T. and Rabusseau, G. (2021). Rademacher random projections with tensor networks. *arXiv preprint arXiv:2110.13970*.
- Schudy, W. and Sviridenko, M. (2012). Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 437–446. SIAM.
- Shi, Y. and Anandkumar, A. (2019). Higher-order Count Sketch: Dimensionality Reduction That Retains Efficient Tensor Operations. arXiv:1901.11261 [cs, stat].
- Stojanac, Z., Suess, D., and Kliesch, M. (2018). On products of Gaussian random variables. arXiv:1711.10516 [math].