Convergence rate of Bayesian supervised tensor modeling with multiway shrinkage priors

Rajarshi Guhaniyogi[1]

Department of Applied Mathematics and Statistics, Baskin School of Engineering,
University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064
email: rguhaniy@ucsc.edu

Abstract

This article studies the convergence rate of the posterior for Bayesian low rank supervised tensor modeling with multiway shrinkage priors. Multiway shrinkage priors constitute a new class of shrinkage prior distributions for tensor parameters in Bayesian low rank supervised tensor modeling to regress a scalar response on a tensor predictor with the primary aim to identify cells in the tensor predictor which are predictive of the scalar response. This novel and computationally efficient framework stems from pressing needs in many applications, including functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI). This article shows that the convergence rate is nearly optimal in terms of in-sample predictive accuracy of the Bayesian supervised low rank tensor model with a multiway shrinkage prior distribution when the number of observations grows. The conditions under which this nearly optimal convergence rate is achieved are seen to be very mild. More importantly, the rate is achieved for an easily computable method, even when the true CP/PARAFAC rank of the tensor coefficient corresponding to the tensor predictor is unknown.

Keywords: Multiway shrinkage prior, Posterior convergence rate, Parafac decomposition, Supervised tensor modeling

1. Introduction

Of late we routinely encounter applications involving predictors having a multidimensional array or tensor structure. For example, in neuroimaging applications, the predictor is often in the form of 3D brain images of an individual consisting of $96\times96\times96$ voxels. Another noteworthy application of tensor predictors arises from brain connectomics, where matrix predictors quantifying connections between different brain regions are used to predict an individual's IQ. The most naive approach assesses association between a tensor predictor and a scalar response by fitting cell by cell independent regression models [11]. Although this approach is widely used for its simplicity, it misses out on important information regarding the way in which multiple cells in a tensor predictor jointly impact a response.

A more sophisticated approach vectorizes the tensor predictor and uses existing high-dimensional regression techniques with the scalar response and the vectorized tensor. Such vectorization fails to capture spatial dependence between tensor cells and suffers in terms of learning the tensor coefficient for small samples. To give an example, in the context of fMRI applications with $96^3 \approx 1$ million image predictors, state-of-the-art Bayesian high-dimensional regression [1, 3] proposes vectorizing the tensor predictor into a vector of dimension 96^3 before regressing it on the scalar response. Gibbs sampling implementation of this high-dimensional regression requires inverting a $96^3 \times 96^3$ matrix, which is infeasible. From the inferential point of view, Bayesian high-dimensional regressions are deemed to be statistically inefficient when the number of predictors exceeds the sample size [2]. In the context of fMRI applications with 96^3 image predictors, this condition demands that consistent estimation of the posterior by vectorizing the tensor predictor is only possible if the sample size exceeds 96^3, an impractical situation in biomedical applications. There is an alternative literature based on functional regression that treats the vectorized tensor as the discretization of a functional predictor [5, 15, 16], though it is not accompanied by proper theoretical guarantee.

It is noteworthy that there is a considerable body of literature both in the theory and application in unsupervised "low rank" tensor modeling on decomposing a tensor into a few factors and identifying the rate at which the distance between true and estimated tensor decays. Here "rank" refers to the PARAFAC or CP-rank [9]. A naive approach to low-rank decomposition of a tensor requires complicated non-convex optimization [9]. Several proposals have been made in the literature to alleviate the computational difficulties caused by the non-convex optimization [6, 12, 18], though they achieve computational efficiency at the expense of a "suboptimal" learning rate. There is a growing literature on Bayesian unsupervised low rank tensor models [4, 8, 20] that constructs a generative model of the tensor decomposition and places a prior probability on low rank decomposed components of the tensor. Though these methods are tailored to model efficiently the low rank decomposition of massive tensors, only a few of them [20] are supported by theoretical results.

Our problem is fundamentally different from estimating unsupervised low rank decomposition of the tensor objects. Rather, we focus on theoretically investigating computationally efficient supervised Bayesian low rank tensor regression models. Our supervised tensor regression framework expresses a regression model in which a tensor coefficient embodies the impact of every cell of the tensor predictor in predicting the scalar response. The prior probability considered on the tensor coefficient is the recently proposed novel Multiway-Dirichlet Generalized Double Pareto Prior (M-DGDP) [7]. It is argued in Guhaniyogi et al. [7] that supervised Bayesian tensor modeling with the M-DGDP prior carefully imparts shrinkage on the tensor coefficient in three different ways: at a global level, at a local level of individual parameters, and by providing shrinkage towards low rank decomposition (in the sense of PARAFAC rank) of the tensor coefficient. A multiway shrinkage prior thus constructed, naturally induces sparsity within and across components in the tensor factorization of the tensor coefficient and exhibits excellent empirical performance in terms of prediction and region selection. Moreover, the M-DGDP prior allows auto tuning of all the hyperparameters with Markov chain Monte Carlo chains showing rapid mixing. Guhaniyogi et al. [7] provide sufficient methodological and applied motivation behind the framework and establish results on posterior consistency for the proposed model.

The major contribution of this article is to offer a stronger theoretical result in estimating the learning rate of the posterior density of the tensor coefficient under mild assumptions. We relax the key assumption in Guhaniyogi et al. [7] that both the tensor predictor coefficient generating the data (also referred to as the true tensor coefficient) and the fitted tensor coefficient have rank R PARAFAC decompositions. In practice, the rank of the true tensor coefficient is never known. Instead, the current article is based upon a more realistic assumption that the rank of the fitted tensor coefficients is merely greater than the rank of the true tensor coefficients. Additionally, Guhaniyogi et al. [7] concentrate exclusively on proving consistency of the posterior distribution, while the present article carefully devises techniques to derive the rate at which the posterior distribution of the tensor coefficient converges to the truth. Roughly speaking, we provide a "near optimal" learning rate of the order of $n^{-1/2}$ up to a $\ln(n)$ factor for the posterior distribution. As a corollary, the Bayes estimate is also shown to have a near optimal convergence rate. These results are considerably stronger than those of Guhaniyogi et al. [7] and also require different proof techniques. Most importantly, to the best of our knowledge, the rate of convergence for coefficients under Bayesian shrinkage priors is not well developed even in the context of ordinary high-dimensional regression. This article provides a posterior concentration rate for shrinkage priors in a tensor regression scenario which is arguably more challenging than ordinary high-dimensional regression with scalar predictors.

Recently, theoretical results on supervised tensor modeling in the frequentist literature [21] have determined the rank of tensor beforehand and therefore, have a different setting from ours. To the best of our knowledge, there is only one prior work [17] that presents a near optimal convergence rate for the posterior distribution of Bayesian supervised tensor modeling with low rank structure on the tensor coefficient. The prior probability considered in Suzuki [17] is the most basic one, which places Gaussian priors on decomposed components of the tensor coefficient and an exponentially decaying prior on the rank. While this prior has optimal theoretical properties, it faces inevitable computational and mixing issues when tensor dimensions are sufficiently large. In contrast, the M-DGDP prior provides a practically useful, computationally efficient posterior with optimal convergence rate.

2. Problem setting

2.1. Notations and definitions

Let $\beta_1 = (\beta_{11}, \ldots, \beta_{1p_1})^\top$ and $\beta_2 = (\beta_{21}, \ldots, \beta_{2p_2})^\top$ be $p_1 \times 1$ and $p_2 \times 1$ vectors, respectively. The vector outer product $\beta_1 \circ \beta_2$ is a $p_1 \times p_2$ array with (i, j)th entry $\beta_{1i}\beta_{2j}$. A D-way outer product between vectors $\beta_j = (\beta_{j1}, \ldots, \beta_{jp_j})^\top$, $j \in \{1, \ldots, D\}$, is a $p_1 \times \cdots \times p_D$ multi-dimensional array denoted by $B = \beta_1 \circ \cdots \circ \beta_D$ with entries $(B)_{i_1,\ldots,i_D} =$

$\beta_{1i_1} \times \cdots \times \beta_{Di_D}$. Define a vec$(B)$ operator as stacking elements of this D-way tensor into a column vector of length $p_1 \times \cdots \times p_D$. From the definition of outer products, it is easy to see that vec$(\beta_1 \circ \cdots \circ \beta_D) = \beta_D \otimes \cdots \otimes \beta_1$. A tensor $B \in \circ_{j=1}^{D} \mathbb{R}^{p_j}$ is known as a D-way tensor.

2.2. *Tensor regression*

Consider data $(y_1, X_1), \ldots, (y_n, X_n)$ where, for each $i \in \{1, \ldots, n\}$, y_i is the scalar response and X_i is the tensor predictor in $\mathbb{R}^{p_1 \times \cdots \times p_D}$. It is assumed that the scalar response y_i is generated from the tensor predictor $X_i = (x_{i,i_1,\ldots,i_D})_{i_1,\ldots,i_D=1}^{p_1,\ldots,p_D}$ following the model

$$y_i = \langle X_i, B_n^0 \rangle + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $B_n^0 = (b^0_{i_1,\ldots,i_D,n})_{i_1,\ldots,i_D=1}^{p_1,\ldots,p_D}$ is the true and unknown tensor coefficient in $\mathbb{R}^{p_1 \times \cdots \times p_D}$. The inner product between two tensors B_n^0 and X_i is defined by $\langle X_i, B_n^0 \rangle = \sum_{i_1,\ldots,i_D} x_{i,i_1,\ldots,i_D} b^0_{i_1,\ldots,i_D,n}$. The errors $\epsilon_1, \ldots, \epsilon_n$ form a random sample from a $\mathcal{N}(0, \sigma^2)$ distribution with mean 0 and variance σ^2.

This article assumes a "low rank" decomposition on the true tensor coefficient B_n^0. Here "rank" refers to the rank of the CP or PARAFAC-decomposition. We say that B_n^0 follows a rank-R_0 PARAFAC decomposition if B_n^0 can be expressed as

$$B_n^0 = \sum_{r=1}^{R_0} \beta_{1,n}^{0(r)} \circ \cdots \circ \beta_{D,n}^{0(r)},$$

for some $\beta_{j,n}^{0(r)} \in \mathbb{R}^{p_j}$, $j \in \{1, \ldots, D\}$ and $r \in \{1, \ldots, R_0\}$, where R_0 is the minimum number to yield such decomposition.

The set

$$\{\beta_{j,n}^{0(r)} \in \mathbb{R}^{p_j} : 1 \le j \le D, \; 1 \le r \le R_0\}$$

is known as the set of tensor margins of B_n^0. Let $B_{j,n}^0 = [\beta_{j,n}^{0(1)} : \cdots : \beta_{j,n}^{0(R_0)}]$ be a $p_j \times R_0$ matrix.

A rank R_0 PARAFAC decomposition of tensor B_n^0 is also presented as $B_n^0 = [[B_{1,n}^0, \ldots, B_{D,n}^0]]$. On a similar note, we fit a tensor regression model to the data $(y_1, X_1), \ldots, (y_n, X_n)$ as follows

$$y_i = \langle X_i, B_n \rangle + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where B_n is the fitted tensor coefficient in $\mathbb{R}^{p_1 \times \cdots \times p_D}$. Similar to the true tensor coefficient, the fitted tensor coefficient B_n is assumed to follow a rank-R PARAFAC decomposition with $B_n = [[B_{1,n}, \ldots, B_{D,n}]]$ and $B_{j,n} = [\beta_{j,n}^{(1)} : \cdots : \beta_{j,n}^{(R)}]$, a $p_j \times R$ matrix. In what follows, we assume for the sake of simplicity that σ^2 is known, and more specifically $\sigma^2 = 1$ without loss of generality. Finally, assume that for all $j \in \{1, \ldots, D\}$ and $i_j \in \{1, \ldots, p_j\}$,

$$\tilde{\beta}_{ji_j,n} = (\beta_{ji_j,n}^{(1)}, \ldots, \beta_{ji_j,n}^{(R)})^\top, \quad \tilde{\beta}_{ji_j,n}^0 = (\beta_{ji_j,n}^{0(1)}, \ldots, \beta_{ji_j,n}^{0(R_0)})^\top.$$

Evidently, the rank of the true tensor (R_0) and the fitted tensor (R) are assumed to be different.

Example 1 (fMRI studies). In neuroscience, often the interest lies in predicting a phenotypic characteristic of an individual based on the functional magnetic resonance imaging (fMRI) obtained from brain scans. fMRI measurements for each individual come in the form of a 3D tensor composed of a large number of cubic cells, known as brain voxels. Typically, a full scale fMRI measurement generates a tensor image consisting of $p_1 \times p_2 \times p_3$ voxels. This acts as a tensor predictor

$$X_i = ((x_{i,i_1,i_2,i_3}))_{i_1,i_2,i_3=1}^{p_1,p_2,p_3}$$

for the ith individual with x_{i,i_1,i_2,i_3} corresponding to the fMRI intensity in the (i_1, i_2, i_3)th voxel. The true and unknown tensor coefficient B_n^0 signifies the weight of each voxel in predicting the phenotypic characteristic. Please refer to Guhaniyogi et al. [7], who investigated tensor regression with fMRI data in detail with the proposed model and prior distributions.

Example 2 (DTI studies). In many neuroscience applications, it is of common interest to build a prediction model of IQ on the brain connectivity. To quantify brain connectivity, important regions of interest (ROI) in the brain are identified and the number of neurons connecting different ROIs is measured from the brain white matter using diffusion tensor imaging (DTI). Let X_i be a matrix of dimension $p \times p$, where p is the number of ROIs. The (t, s)th entry in X_i, denoted by $x_{i,t,s}$, is given by the number of neurons connecting ROI s to ROI t. The goal is to predict IQ (a scalar response) based on the $p \times p$ connectivity matrix. Here the true tensor B_n^0 quantifies the effect of neuron connectivity between two ROIs in predicting the IQ. For more details, please see Guhaniyogi et al. [7], who present a detailed analysis of the tensor regression model with the M-DGDP prior on the tensor coefficients for the DTI data.

The goal in these examples is to estimate the unknown tensor coefficient B_n^0 and to facilitate the accurate prediction of y based on X. Note that estimation of a higher-dimensional tensor B_n^0 in Example 1 is much more complex than estimating a matrix. Zhou et al. [21] proposed a theoretically optimal frequentist procedure to estimate B_n^0 under the assumption that R_0 is known.

Instead of convex regularized point estimation, this article provides a Bayesian procedure to estimate the posterior distribution of B_n. It will be shown in due course that the posterior predictive loss (defined in Section 2.4) of our procedure decays at the "near" optimal rate to 0 under weak assumptions. Moreover, the posterior is easily computable with standard Markov chain Monte Carlo updates for all the parameters.

2.3. *Prior and posterior distributions of B_n*

This section discusses the choice of prior and the induced posterior distribution on B_n. Note that there has been a growing interest, in high-dimensional regression with vector predictors, in choosing priors on predictor coefficients which shrink small coefficients towards zero while maintaining minimum shrinkage for large coefficients. Many of these priors design shrinkage through a global parameter and a set of local parameters. The global parameter imposes shrinkage globally while local parameters carefully balance shrinkage for large and small coefficients [14].

The literature on the vector shrinkage priors provides an excellent starting point for studying multiway shrinkage prior on tensor coefficient B_n, though the latter presents a lot more hurdles. Assuming that B_n admits a rank-R PARAFAC decomposition, proposing a prior on B_n is equivalent to specifying priors over tensor margins $\beta_{j,n}^{(r)}$. Given that every cell coefficient in B_n is a nonlinear function of the tensor margins, care should be taken while imposing prior shrinkage on them. To this end, Guhaniyogi et al. [7] characterize multiple restrictions on putting prior distributions on $\beta_{j,n}^{(r)}$'s to facilitate desirable shrinkage for the posterior distribution of $b_{i_1,\ldots,i_D,n}$ and propose one multiway shrinkage prior satisfying all the restrictions.

This article provides a theoretical analysis of the proposed multiway shrinkage prior over B_n deemed the multiway Dirichlet generalized double Pareto (M-DGDP) prior [7]. The M-DGDP prior induces shrinkage across components in an exchangeable way, setting $\tau_r = \phi_r \tau$ as the global scale for component $r \in \{1, \ldots, R\}$, with $\tau \sim \mathcal{GA}(a_\tau, b_\tau)$ and $\Phi = (\phi_1, \ldots, \phi_R) \sim \mathcal{DIR}(\alpha_1, \ldots, \alpha_R)$. The hierarchical margin-level prior is given by

$$\beta_{j,n}^{(r)} \sim \mathcal{N}[0, (\phi_r \tau) W_{jr}], \quad w_{jr,k} \sim \mathcal{E}(\lambda_{jr}^2/2), \quad \lambda_{jr} \sim \mathcal{G}(a_\lambda, b_\lambda). \tag{2}$$

Additional flexibility in estimating $\{\beta_{j,n}^{(r)} : 1 \leq j \leq D\}$ is accommodated by modeling heterogeneity within margins via element-specific scaling $w_{jr,k}$. Above, $W_{jr} = \operatorname{diag}(w_{jr,1}, \ldots, w_{jr,p_j})$ are local (margin and component-specific) scale parameters for each margin $j \in \{1, \ldots, D\}$ and every component $r \in \{1, \ldots, R\}$. A common rate parameter λ_{jr} encourages sharing of information between the marginal elements. Collapsing over the element-specific scales, one has, for all $k \in \{1, \ldots, p_j\}$,

$$\beta_{j,k,n}^{(r)} \mid \lambda_{jr}, \phi_r, \tau \overset{\text{iid}}{\sim} \mathcal{DE}(\lambda_{jr}/\sqrt{\phi_r \tau}).$$

Prior (2) leads to a generalized double pareto (GDP) shrinkage prior having the form of an adaptive Lasso penalty on the individual margin coefficients.

Let the likelihood of (1) be denoted by $f(y_{1:n} \mid B_n, X_{1:n})$ so that

$$f(y_{1:n} \mid B_n, X_{1:n}) \propto \exp\Big\{-\sum_{i=1}^{n} (y_i - \langle X_i, B_n \rangle)^2/2\Big\}.$$

Denoting the prior distribution of B_n by $\pi(B_n)$, the posterior distribution of B_n is given by

$$\Pi(B_n \in \mathcal{B}_n \mid y_{1:n}, X_{1:n}) = \frac{\int_{\mathcal{B}_n} f(y_{1:n} \mid B_n, X_{1:n}) \pi(B_n) dB_n}{\int f(y_{1:n} \mid B_n, X_{1:n}) \pi(B_n) dB_n},$$

where $\mathcal{B}_n$ is a subset of $\mathbb{R}^{p_1\times\cdots\times p_D}$. It is easy to see that the conditional posteriors for all parameters are in closed form. Therefore, Gibbs sampling [7] can readily be employed to estimate the marginal posterior distribution of B_n.

2.4. Convergence rate analysis

This section presents a convergence rate analysis for the posterior distribution of B_n. We first start by defining a few quantities. Let the L_p, L_∞ and empirical distance between two tensors B_n and B_n' be given by

$$
\begin{aligned}
\|B_n - B_n'\|_p &= \left(\sum_{i_1,\ldots,i_D=1}^{p_1,\ldots,p_D} |b_{i_1,\ldots,i_D,n} - b'_{i_1,\ldots,i_D,n}|^p \right)^{1/p}, \\
\|B_n - B_n'\|_\infty &= \max_{i_1,\ldots,i_D} |b_{i_1,\ldots,i_D,n} - b'_{i_1,\ldots,i_D,n}|, \\
\|B_n - B_n'\|_n &= \left(\frac{1}{n} \sum_{i=1}^{n} \langle X_i, B_n - B_n' \rangle^2 \right)^{1/2},
\end{aligned}
$$

respectively, where $B_n = ((b_{i_1,\ldots,i_D,n}))_{i_1,\ldots,i_D=1}^{p_1,\ldots,p_D}$ and $B_n' = ((b'_{i_1,\ldots,i_D,n}))_{i_1,\ldots,i_D=1}^{p_1,\ldots,p_D}$.

It is argued by van der Vaart & Van Zanten [19] that the predictive accuracy of (1) can readily be assessed by investigating the rate of convergence with respect to n of the risk

$$
\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{B_n^0} \int \mathrm{KL}\{f(y_i \mid X_i, B_n), f(y_i \mid X_i, B_n^0)\} \pi(B_n \mid y_{1:n}, X_{1:n}).
$$

The KL divergence of two normal densities $f(y_i \mid X_i, B_n)$ and $f(y_i \mid X_i, B_n^0)$ with means $\langle X_i, B_n\rangle$ and $\langle X_i, B_n^0\rangle$ and variance 1 is equal to $(\langle X_i, B_n - B_n^0\rangle)^2/2$. Therefore, the risk reduces to

$$
\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{B_n^0} \int \mathrm{KL}\{f(y_i \mid X_i, B_n), f(y_i \mid X_i, B_n^0)\} \pi(B_n \mid y_{1:n}, X_{1:n}) = \frac{1}{2}\, \mathrm{E}_{B_n^0} \int \|B_n - B_n^0\|_n^2 \pi(B_n \mid y_{1:n}, X_{1:n}). \tag{3}
$$

If the above risk is bounded by ϵ_n^2 for some $\epsilon_n \to 0$, then by applying Jensen's inequality one can see that the Bayes estimator satisfies $\mathrm{E}_{B_n^0}\|\mathrm{E}(B_n \mid y_{1:n}, X_{1:n}) - B_n^0\|_n^2 \le \epsilon_n^2$. Thus the Bayes estimator converges at the same rate ϵ_n to the true tensor coefficient B_n^0. In the rest of the article we focus on obtaining ϵ_n for the supervised Bayesian tensor modeling.

One of the key quantities in proving posterior convergence rate results is the concentration of the prior distribution. The prior concentration can be quantified by $\mathcal{E}$, defined, for each $\delta > 0$, by

$$
\mathcal{E}(\delta) = -\ln\{\pi(B_n : \|B_n - B_n^0\|_n < \delta)\}.
$$

For the posterior to have an "optimal" rate of convergence, one expects the prior to put considerable mass around B_n^0. Again, B_n^0 is unknown, and so it is not desirable to have lots of prior mass around a point or a few points. Rather the prior mass should be spread judiciously, taking into account the wide range of possible B_n^0 values. Lemma 1 below presents an upper bound on the prior concentration of the M-DGDP prior.

Lemma 1. *Let* $C^{-1} = \Gamma(R\alpha)\{\Gamma(\alpha)\}^{-R}\{\Gamma(\alpha + a_\lambda D/2)\}^{R}\{\Gamma(R\alpha + Ra_\lambda D/2)\}^{-1} \exp(-b_\tau) b_\tau^{a_\tau}\, b_\lambda^{DRa_\lambda}(b_\tau + a_\lambda RD/2)^{-1}$. *Further, assume that* $\Delta_{i_1,\ldots,i_D}$ *is a positive root of the equations given, for all* $i_j \in \{1,\ldots,p_j\}$ *and* $j \in \{1,\ldots,D\}$, *by*

$$
x(x + \|\tilde\beta^0_{2i_2,n}\|)\cdots(x + \|\tilde\beta^0_{Di_D,n}\|) + \|\tilde\beta^0_{1i_1,n}\| x(x + \|\tilde\beta^0_{2i_2,n}\|)\cdots(x + \|\tilde\beta^0_{Di_D,n}\|) + \cdots + x\|\tilde\beta^0_{2i_2,n}\|\cdots\|\tilde\beta^0_{D-1i_{D-1},n}\| - \delta = 0, \tag{4}
$$

and $\Delta = \min_{i_1,\ldots,i_D}\Delta_{i_1,\ldots,i_D}$. *Then, for* $R > R_0$,

$$
\begin{aligned}
\mathcal{E}(\delta) \le \left(R\sum_{j=1}^{D} p_j \right) \ln\{(2\pi R)^{1/2}/(2\Delta)\} - \ln(C) + R\sum_{j=1}^{D} \ln\{\Gamma(a_\lambda)/\Gamma(a_\lambda + p_j)\} \\
+ \sum_{j=1}^{D}\sum_{r=1}^{R_0} (a_\lambda + p_j) \ln\left[b_\lambda + \sum_{i_j=1}^{p_j} \{(\beta^{0(r)}_{ji_j,n})^2 + 2\Delta^2\}^{1/2} \right] + (R - R_0)\sum_{j=1}^{D} (a_\lambda + p_j)\ln(b_\lambda + p_j 2^{1/2}\Delta).
\end{aligned}
$$

Evidently, $\mathcal{E}(\delta_1) \le \mathcal{E}(\delta_2)$ for $0 < \delta_2 < \delta_1$.

2.5. Analysis of the in-sample predictive accuracy

This section discusses bounds on the *in-sample predictive accuracy* of (1). Assuming n sample points (y_1, X_1), ..., (y_n, X_n) with X_i's fixed, Theorem 2 provides the rate at which the posterior of B_n converges to the data generating tensor coefficient B_n^0 under the metric (3). The proof of the theorem is given in the Appendix.

Theorem 2. *Assume that (i) $||X_i||_2 \leq 1$; (ii) there exists a constant M such that $||\tilde{\beta}^0_{ji_j,n}||_2 < M$, for all $i_j \in \{1, \dots, p_j\}$, $j \in \{1, \dots, D\}$, $r \in \{1, \dots, R_0\}$; (iii) there exists another constant L such that $p_j < L$ for all $j \in \{1, \dots, D\}$; (iv) $a_\lambda > R(p_1 + \cdots + p_D)$. Under these assumptions, the in sample predictive accuracy is upper bounded by a quantity given below*

$$\mathrm{E}_{B_n^0} \int ||B_n - B_n^0||_n^2 \pi(B_n \mid y_{1:n}, X_{1:n}) \leq AH_n/n,$$

where $H_n = o\{\ln(n)^d\}$ and A are constants depending on $n, D, L, a_\lambda, b_\lambda, a_\tau, M, R$ for any d.

By the discussion provided after Eq. (3), it is evident that Theorem 2 proposes a much stronger result than the mere convergence rate of the posterior mean estimator. In the frequentist literature, it is usually a common practice to assume a variant of strong convexity such as restricted strong convexity or restricted eigenvalue property [13] to derive fast convergence rates for sparse estimators. However, Theorem 2 does not require assuming strong convexity in the design. Similar to ours, the convergence rate result presented in Suzuki [17] also avoids assuming any strong convexity. However, the strongest point of our analysis remains in deriving convergence rate results for an easily implementable multiway shrinkage prior for large tensors.

3. Discussion

This article investigates the convergence rate of the posterior distribution for the supervised Bayesian low rank tensor model proposed in Guhaniyogi et al. [7]. The convergence rate is found to be "near optimal", is obtained under very mild conditions and without any assumption of strong convexity. In contrast with the frequentist tensor regression, our analysis does not assume that the true rank R_0 is known. Most importantly, the bound on the predictive accuracy is achieved for a novel multiway shrinkage prior distribution that leads to an easily computable posterior.

Several future directions of research emerge from this article. Note that this article assumes an upper bound on the tensor dimensions. Sometimes it might be of interest to assess the rate of convergence when the dimension of the tensor grows with the sample size. To this end, one assumes p_j as a function of n, say $p_{j,n}$, and investigates the convergence rate. Similarly, one might also allow $||\tilde{\beta}^0_{ji_j,n}||_2$ to vary slowly as a function of n and investigate the change in the rate of convergence. Another interesting future direction constitutes extending this theoretical set up to a more general low rank supervised tensor model with Tucker decomposition of the tensor coefficient.

Appendix

Proof of Lemma 1. First note that for all $r \in \{R_0 + 1, \dots, R\}$,

$$\begin{aligned}
|b_{i_1,\dots,i_D,n} - b^0_{i_1,\dots,i_D,n}| &= \left| \sum_{r=1}^{R} \beta^{(r)}_{1i_1,n} \cdots \beta^{(r)}_{Di_D,n} - \sum_{r=1}^{R} \beta^{0(r)}_{1i_1,n} \cdots \beta^{0(r)}_{Di_D,n} \right| \\
&= \left| \sum_{r=1}^{R} \left\{ (\beta^{(r)}_{1i_1,n} - \beta^{0(r)}_{1i_1,n}) \prod_{j \neq 1} \beta^{(r)}_{ji_j,n} + \cdots + (\beta^{(r)}_{Di_D,n} - \beta^{0(r)}_{Di_D,n}) \prod_{j \neq D} \beta^{0(r)}_{ji_j,n} \right\} \right| \\
&\leq ||\tilde{\beta}_{1i_1,n} - \tilde{\beta}^0_{1i_1,n}||_2 \prod_{j \neq 1} ||\tilde{\beta}_{ji_j,n}||_2 + \cdots + ||\tilde{\beta}_{Di_D,n} - \tilde{\beta}^0_{Di_D,n}||_2 \prod_{j \neq D} ||\tilde{\beta}^0_{ji_j,n}||_2,
\end{aligned}$$

where $\beta^{0(r)}_{ji_j,n} = 0$. Note that (4) can be written as $\mathrm{g}_{i_1,\dots,i_D}(x) = 0$, where

$$\mathrm{g}_{i_1,\dots,i_D}(x) = a_{D,i_1,\dots,i_D} x^D + \cdots + a_{1,i_1,\dots,i_D} x - a_{0,i_1,\dots,i_D}$$

and the $a_{i,i_1,\ldots,i_D}$'s are suitably chosen to match the coefficient of x^i in (4). By Cauchy's bound on the roots of polynomials, Eq. (4) has only one positive root, namely the real $\Delta_{i_1,\ldots,i_D}$ that satisfies $\Delta_{i_1,\ldots,i_D} \leq 1 + \max_{i=0,\ldots,D}|a_{i,i_1,\ldots,i_D}|$, for all $i_1, \ldots, i_D$. From (4), the fact that $||\tilde{\beta}_{ji_j,n} - \tilde{\beta}^0_{ji_j,n}|| < \Delta$ for all $i_j \in \{1, \ldots, p_j\}$ and $j \in \{1, \ldots, D\}$ implies

$$|b_{i_1,\ldots,i_D,n} - b^0_{i_1,\ldots,i_D,n}| \leq \mathrm{g}_{i_1,\ldots,i_D}(\Delta) + \delta \leq \mathrm{g}_{i_1,\ldots,i_D}(\Delta_{i_1,\ldots,i_D}) + \delta = \delta,$$

which leads to $||B_n - B^0_n||_\infty < \delta$. Hence

$$\Pi(B_n : ||B_n - B^0_n||_n < \delta) \geq \Pi(B_n : ||B_n - B^0_n||_\infty < \delta) \geq \Pi(\forall_{j\in\{1,\ldots,D\}}\ \forall_{i_j\in\{1,\ldots,p_j\}}\ ||\tilde{\beta}_{ji_j,n} - \tilde{\beta}^0_{ji_j,n}||_2 < \Delta).$$

Therefore, it is enough to bound the right-hand side from below. One has

$$\begin{aligned}
&\Pi\left(\forall_{j\in\{1,\ldots,D\}}\ \forall_{i_j\in\{1,\ldots,p_j\}}\ ||\tilde{\beta}_{ji_j,n} - \tilde{\beta}^0_{ji_j,n}||_2 < \Delta \mid \{\phi_r\}, \tau, \{W_{jr}\}\right) \\
&\quad = \prod_{j=1}^{D}\prod_{i_j=1}^{p_j}\left[\exp\left\{-\sum_{r=1}^{R}(\beta^{0(r)}_{ji_j,n})^2/(2w_{jr,i_j}\phi_r\tau)\right\}\Pi\left(||\tilde{\beta}_{ji_j,n}|| < \Delta/2 \mid \{\phi_r\}, \tau, \{W_{jr}\}\right)\right] \\
&\quad \geq \prod_{j=1}^{D}\prod_{i_j=1}^{p_j}\left[\exp\left\{-\sum_{r=1}^{R}(\beta^{0(r)}_{ji_j,n})^2/(2w_{jr,i_j}\phi_r\tau)\right\}\prod_{r=1}^{R}\left[\exp\{-\Delta^2/(\phi_r\tau w_{jr,i_j})\}(2\Delta)/(2\pi R\phi_r\tau w_{jr,i_j})^{1/2}\right]\right] \\
&\quad \geq \prod_{j=1}^{D}\prod_{i_j=1}^{p_j}\prod_{r=1}^{R}\left[(2\Delta)/(2\pi R\phi_r\tau w_{jr,i_j})^{1/2}\exp\left[-\{\Delta^2 + (\beta^{0(r)}_{ji_j,n})^2/2\}/(\phi_r\tau w_{jr,i_j})\right]\right],
\end{aligned}$$

where Step 2 follows from Anderson's lemma. Integrating out the w_{jr,i_j}'s, we obtain

$$\begin{aligned}
&\Pi\left(\forall_{j\in\{1,\ldots,D\}}\ \forall_{i_j\in\{1,\ldots,p_j\}}\ ||\tilde{\beta}_{ji_j,n} - \tilde{\beta}^0_{ji_j,n}|| < \Delta \mid \tau, \{\phi_r\}, \{\lambda_{jr}\}\right) \\
&\qquad \geq \prod_{r=1}^{R}\prod_{j=1}^{D}\left[\{(2\Delta\lambda_{jr})/(R\phi_r\tau)^{1/2}\}^{p_j}\exp\left[-\lambda_{jr}\sum_{i_j=1}^{p_j}\{(\beta^{0(r)}_{ji_j,n})^2 + 2\Delta^2\}^{1/2}/(\phi_r\tau)^{1/2}\right]\right].
\end{aligned}$$

Integrating out the λ_{jr}'s, we then get

$$\begin{aligned}
&\Pi\left(\forall_{j\in\{1,\ldots,D\}}\ \forall_{i_j\in\{1,\ldots,p_j\}}\ ||\tilde{\beta}_{ji_j,n} - \tilde{\beta}^0_{ji_j,n}|| < \Delta \mid \tau, \{\phi_r\}\right) \\
&\geq \prod_{r=1}^{R}\prod_{j=1}^{D}\left[\{(2\Delta)/(R\phi_r\tau)^{1/2}\}^{p_j}\Gamma(a_\lambda + p_j)/\left[b_\lambda + \sum_{i_j=1}^{p_j}\{(\beta^{0(r)}_{ji_j,n})^2 + 2\Delta^2\}^{1/2}(\phi_r\tau)^{-1/2}\right]^{a_\lambda+p_j}\right]\{b_\lambda^{a_\lambda}/\Gamma(a_\lambda)\}^{RD} \\
&\geq \prod_{r=1}^{R}\prod_{j=1}^{D}\left[\{(2\Delta)/(R\phi_r\tau)^{1/2}\}^{p_j}\{b_\lambda^{a_\lambda}/\Gamma(a_\lambda)\}\frac{\Gamma(a_\lambda + p_j)(\phi_r\tau)^{(a_\lambda+p_j)/2}\mathbf{1}\{\tau \in (0,1)\}}{\left[b_\lambda + \sum_{i_j=1}^{p_j}\{(\beta^{0(r)}_{ji_j,n})^2 + 2\Delta^2\}^{1/2}\right]^{a_\lambda+p_j}}\right]
\end{aligned}$$

Finally, integrating out τ, leads to

$$\begin{aligned}
&\Pi(\forall_{j\in\{1,\ldots,D\}}\ \forall_{i_j\in\{1,\ldots,p_j\}}\ ||\tilde{\beta}_{ji_j,n} - \tilde{\beta}^0_{ji_j,n}|| < \Delta) \\
&\qquad \geq \prod_{j=1}^{D}\{\Gamma(a_\lambda + p_j)/\Gamma(a_\lambda)\}^R\prod_{j=1}^{D}\prod_{r=1}^{R}\left[b_\lambda + \sum_{i_j=1}^{p_j}\{(\beta^{0(r)}_{ji_j,n})^2 + 2\Delta^2\}^{1/2}\right]^{-a_\lambda-p_j}\{2\Delta/(2\pi R)^{1/2}\}^{R\sum_{j=1}^{D}p_j}C^{-1}.
\end{aligned}$$

This completes the proof of Lemma 1. □

Next we prove two lemmas which will be central in proving Theorem 2. In what follows,

$$\mathcal{F}_s = \left\{B_n = [[B_{1,n}, \ldots, B_{D,n}]] : \max_{i_j=1:p_j, j=1:D}||\tilde{\beta}_{ji_j,n}||_2 \leq M_s^{1/2}\right\}.$$

Lemma 3. *Assume that $\mathcal{D}(g, \mathcal{G}, \|\cdot\|)$ is the minimum number of disjoint balls of radius g under $\|\cdot\|$ norm to completely cover $\mathcal{G}$, also known as the packing number of $\mathcal{G}$. Then* $\ln\{\mathcal{D}(n^{1/2}\delta, \mathcal{F}_s, n^{1/2}\|\cdot\|_n)\} \leq \{R(p_1 + \cdots + p_D)\} \ln\{(5DM_s^{D/2})/\delta\}$.

Proof. Note that

$$\begin{aligned}
\|B_n - B_n'\|_2^2 &= \sum_{i_1,\dots i_D=1}^{p_1,\dots,p_D} \left\{ \sum_{r=1}^{R} \left(\beta_{1i_1,n}^{(r)} \cdots \beta_{Di_D,n}^{(r)} - \beta_{1i_1,n}^{(r)'} \cdots \beta_{Di_D,n}^{(r)'} \right) \right\}^2 \\
&= \sum_{i_1,\dots i_D=1}^{p_1,\dots,p_D} \left[\sum_{r=1}^{R} \left\{ (\beta_{1i_1,n}^{(r)} - \beta_{1i_1,n}^{(r)'}) \prod_{j\neq 1} \beta_{ji_j,n}^{(r)} + \cdots + (\beta_{Di_D,n}^{(r)} - \beta_{Di_D,n}^{(r)'}) \prod_{j\neq D} \beta_{ji_j,n}^{(r)'} \right\} \right]^2 \\
&\leq D \sum_{i_1,\dots i_D=1}^{p_1,\dots,p_D} \left\{ \|\beta_{1i_1,n} - \beta_{1i_1,n}'\|_2^2 \prod_{j\neq 1} \|\beta_{ji_j,n}\|_2^2 + \cdots + \prod_{j\neq D} \|\beta_{ji_j,n}'\|_2^2 \|\beta_{Di_D,n} - \beta_{Di_D,n}'\|_2^2 \right\} \\
&\leq DM_s^{D-1} \sum_{i_1,\dots i_D=1}^{p_1,\dots,p_D} (\|\beta_{1i_1,n} - \beta_{1i_1,n}'\|_2^2 + \cdots + \|\beta_{Di_D,n} - \beta_{Di_D,n}'\|_2^2).
\end{aligned}$$

From the above results we get

$$\begin{aligned}
\ln\{\mathcal{D}(n^{1/2}\delta, \mathcal{F}_s, n^{1/2}\|\cdot\|_n)\} &\leq \ln\{\mathcal{D}(\delta, \mathcal{F}_s, \|\cdot\|_2)\} \\
&\leq \ln\{\mathcal{D}(\delta/(DM_s^{(D-1)/2}), \mathcal{B}_{R\sum_{j=1}^{D} p_j}(M_s^{1/2}), \|\cdot\|_2)\} \\
&\leq \Big(R \sum_{j=1}^{D} p_j\Big) \ln[\{4 + \delta/(DM_s^{D/2})\}/\{\delta/(DM_s^{D/2})\}] \leq \Big(R \sum_{j=1}^{D} p_j\Big) \ln\{(5DM_s^{D/2})/\delta\}.
\end{aligned}$$

This completes the proof of Lemma 3. □

Lemma 4. *The prior probability of $\mathcal{F}_s^{\mathsf{C}}$ is bounded above by* $(RD)^{a_\lambda+1} b_\lambda^{a_\lambda} M_s^{-a_\lambda/2} b_\tau^{-a_\lambda/2} \{\Gamma(a_\tau + a_\lambda/2)/\Gamma(a_\tau)\}$.

Proof. We have

$$\begin{aligned}
\Pi(\mathcal{F}_s^{\mathsf{C}}) &= \Pi(\{B_n = [[B_{1,n}, \dots, B_{D,n}]] : \max_{i_j=1:p_{j,n}, j=1:D} \|\tilde{\beta}_{ji_j,n}\|_2 > M_s^{1/2}\}) \\
&\leq \Pi\Big(\sum_{r=1}^{R} \sum_{j=1}^{D} \sum_{i_j=1}^{p_{j,n}} |\beta_{ji_j,n}^{(r)}| > M_s^{1/2} \Big) \\
&= \sum_{r=1}^{R} \sum_{j=1}^{D} \mathrm{E}\Big[\Pi \Big\{ \sum_{i_j=1}^{p_{j,n}} |\beta_{ji_j,n}^{(r)}| > M_s^{1/2}/(RD) | \{\lambda_{jr}\}, \{\phi_r\}, \tau \Big\} \Big] \\
&\leq \sum_{r=1}^{R} \sum_{j=1}^{D} \mathrm{E}[\exp[-(M_s^{1/2}\lambda_{jr})/\{16RD(\phi_r\tau)^{1/2}\}] | \{\lambda_{jr}\}, \{\phi_r\}, \tau].
\end{aligned}$$

The above uses the fact that $\sum_{i_j=1}^{p_j} |\beta_{ji_j,n}^{(r)}| \sim \mathcal{GA}(p_j, \lambda_{jr}/(\phi_r\tau)^{1/2})$ given $\lambda_{jr}, \phi_r, \tau$, and Lemma 1 in [10]. Further,

$$\begin{aligned}
&\sum_{j=1}^{D} \sum_{r=1}^{R} \mathrm{E}[\exp[-M_s^{1/2}\lambda_{jr}/\{16RD(\phi_r\tau)^{1/2}\}] | \{\lambda_{jr}\}, \{\phi_r\}, \tau] \\
&\quad = \sum_{j=1}^{D} \sum_{r=1}^{R} \mathrm{E}[b_\lambda^{a_\lambda}/[b_\lambda + \{1/(16RD)\}\{M_s/(\phi_r\tau)\}^{1/2}]^{a_\lambda} | \{\phi_r\}, \tau] \\
&\quad \leq \sum_{j=1}^{D} \sum_{r=1}^{R} \mathrm{E}[b_\lambda^{a_\lambda}/[b_\lambda + \{1/(16RD)\}(M_s/\tau)^{1/2}]^{a_\lambda} | \tau] \leq (RD)^{a_\lambda+1} b_\lambda^{a_\lambda} M_s^{-a_\lambda/2} \mathrm{E}(\tau^{a_\lambda/2}) \\
&\quad = (RD)^{a_\lambda+1} b_\lambda^{a_\lambda} M_s^{-a_\lambda/2} b_\tau^{-a_\lambda/2} \{\Gamma(a_\tau + a_\lambda/2)/\Gamma(a_\tau)\}.
\end{aligned}$$

This completes the proof of Lemma 4. □

Proof of Theorem 2 For a test function ϕ_s, event $\mathcal{A}_s$ and set of tensors $\mathcal{F}_s$ (all of them depending on s), we have

$$\begin{aligned} &\mathrm{E}\left\{\int \|B_n - B_n^0\|_n^2 \Pi(B_n \mid y_{1:n}, X_{1:n})\right\} = \mathrm{E}\left\{32\epsilon_n^2 \int_{s>0} s\Pi(\|B_n - B_n^0\|_n > 4s\epsilon_n \mid y_{1:n}, X_{1:n})\right\} \\ &\le 32\epsilon_n^2 \int_{s>0} s\,(A_s + B_s + C_s + D_s)\,ds = 32\epsilon_n^2 + 32\epsilon_n^2 \int_{s>1} s\,(A_s + B_s + C_s + D_s)\,ds, \end{aligned}$$

where $A_s = \mathrm{E}(\phi_s)$, $T_s = \Pr(\mathcal{A}_s^{\complement})$, $C_s = \mathrm{E}\{(1-\phi_s)\mathbf{1}_{\mathcal{A}_s}\Pi(B_n \in \mathcal{F}_s^{\complement} \mid y_{1:n}, X_{1:n})\}$ and

$$D_s = \mathrm{E}\{(1-\phi_s)\mathbf{1}_{\mathcal{A}_s}\Pi(B_n \in \mathcal{F}_s : \|B_n - B_n^0\|_n > 4\epsilon_n s \mid y_{1:n}, X_{1:n})\}.$$

a) Bounding A_s: For any arbitrary $s' > 0$, define $\mathcal{H}_{j,s'} = \{B_n \in \mathcal{F}_s : js' \le n^{1/2}\|B_n - B_n^0\|_n \le (j+1)s'\}$. Let $C_{j,s'} \subset \mathcal{H}_{j,s'}$ be the maximum cardinality set such that $B_n, B_n' \in C_{j,s'} \Rightarrow n^{1/2}\|B_n - B_n'\|_n \ge js'/2$. The cardinality of $C_{j,s'}$ is $\mathcal{D}(js'/2, \mathcal{H}_{j,s'}, n^{1/2}\|\cdot\|_n)$. Using similar arguments as in [17, 19], there exists a test function $\phi_{j,s}$ such that

$$\begin{aligned} &\mathrm{E}_{B_n^0}(\phi_{j,s}) \le 9\exp(-s'^2/8), \\ \sup_{B_n \in \mathcal{H}_{j,s'}, n^{1/2}\|B_n - B_n^0\|_n \ge s'} &\mathrm{E}_{B_n}(1-\phi_{j,s}) \le \mathcal{D}(s'/2, \mathcal{F}_s, n^{1/2}\|\cdot\|_n)\exp(-j^2 s'^2/8). \end{aligned}$$

Construct $\phi_s = \max_{j\ge1}\phi_{j,s}$ and take $s' = 4n^{1/2}\epsilon_n s$. Then the above equations lead to

$$\mathrm{E}_{B_n^0}(\phi_s) \le 9\exp(-2n\epsilon_n^2 s^2), \tag{5}$$

$$\sup_{\|B_n - B_n^0\|_n \ge 4\epsilon_n s} \mathrm{E}_{B_n}(1-\phi_s) \le \mathcal{D}(2n^{1/2}\epsilon_n s, \mathcal{F}_s, n^{1/2}\|\cdot\|_n)\exp(-2s^2 n\epsilon_n^2). \tag{6}$$

b) Bounding T_s: From the proof of [19], there exists an event $\mathcal{A}_s$ such that

$$\Pr_{B_n^0}(\mathcal{A}_s^{\complement}) \le \exp(-n\epsilon_n^2 s^2/8), \tag{7}$$

and on the event $\mathcal{A}_s$, the following inequality holds:

$$\int \{f(y_{1:n} \mid B_n)/f(y_{1:n} \mid B_n^0)\}\Pi(B_n) \ge \exp(-n\epsilon_n^2 s^2)\Pi(B_n : \|B_n - B_n^0\|_n < \epsilon_n s). \tag{8}$$

c) Bounding C_s: Note that for any event $\mathcal{B}$ on $\mathcal{A}_s$,

$$\begin{aligned} \Pi(\mathcal{B} \mid y_{1:n}) &= \left\{\int_{\mathcal{B}} f(y_{1:n} \mid B_n)/f(y_{1:n} \mid B_n^0)\Pi(B_n)\right\} \Big/ \left\{\int f(y_{1:n} \mid B_n)/f(y_{1:n} \mid B_n^0)\Pi(B_n)\right\} \\ &\le \exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\} \int_{\mathcal{B}} f(y_{1:n} \mid B_n)/f(y_{1:n} \mid B_n^0)\Pi(B_n), \end{aligned}$$

which follows from (8). Thus

$$\begin{aligned} C_s &= \mathrm{E}_{B_n^0}\{\Pi(\mathcal{F}_s^{\complement} \mid y_{1:n})\mathbf{1}_{\mathcal{A}_s}(1-\phi_s)\} \\ &\le \exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\} \int_{\mathcal{F}_s^{\complement}} \int_{\mathbb{R}^n} f(y_{1:n} \mid B_n)(1-\phi_s)\Pi(B_n) \\ &\le \exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\}\Pi(\mathcal{F}_s^{\complement}) \le \exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\}(RD)^{a_\lambda+1} b_\lambda^{a_\lambda}\mathrm{E}(\tau^{a_\lambda/2})M_s^{-a_\lambda/2} \end{aligned}$$

d) Bounding D_s:

$$\begin{aligned} D_s &= \mathrm{E}_{B_n^0}\left\{(1-\phi_s)\mathbf{1}_{\mathcal{A}_s}\Pi(B_n \in \mathcal{F}_s : \|B_n - B_n^0\|_n > 4\epsilon_n s \mid y_{1:n})\right\} \\ &\leq \exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\} \sum_{j=1}^{\infty} \int_{\mathbb{R}^n} \int_{B_n \in \mathcal{H}_{j,s}} (1-\phi_s) f(y_{1:n} \mid B_n)\Pi(B_n) \\ &\leq \exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\}\mathcal{D}(2n^{1/2}\epsilon_n s, \mathcal{F}_s, n^{1/2}\|\cdot\|_n)\exp(-2s^2 n\epsilon_n^2), \end{aligned}$$

where the last line follows from (6).

Choose $M_s = \exp(nc\epsilon_n^2 s^2)$, where $2/\{R(p_1 + \cdots + p_D)\} \geq c \geq 2/a_\lambda$. Then

$$\int_{s>1} sA_s ds \leq \int_{s>1} s\exp(-2n\epsilon_n^2 s^2)ds = \int_{s>1} \exp(-2n\epsilon_n^2 s)ds \leq 1/(2n\epsilon_n^2). \tag{9}$$

Similarly, from (7)

$$\int_{s>1} sT_s ds \leq \int_{s>1} s\mathrm{Pr}_{B_n^0}(\mathcal{A}_s^{\complement})ds \leq 8/(n\epsilon_n^2). \tag{10}$$

Furthermore, one has

$$\begin{aligned} \int_{s>1} sC_s ds &\leq \int_{s>1} s\exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\}(RD)^{a_\lambda+1} b_\lambda^{a_\lambda}\mathrm{E}(\tau^{a_\lambda/2})M_s^{-a_\lambda/2} \\ &= \int_{s>1} s\exp\{-n\epsilon_n^2 s^2(a_\lambda c/2 - 1)\}(RD)^{a_\lambda+1} b_\lambda^{a_\lambda}\Gamma(a_\tau + a_\lambda)\exp\{\mathcal{E}(s\epsilon_n)\}\{b_\tau^{a_\lambda}\Gamma(a_\tau)\}^{-1} \\ &= \{5D/(2n^{1/2}\epsilon_n)\}^{R\sum_{j=1}^D p_j} C^{-1}(2\pi R)^{R\sum_{j=1}^D p_j/2} \prod_{j=1}^{D}\{\Gamma(a_\lambda)/\Gamma(a_\lambda + p_j)\}^R \\ &\quad \times \int_{s>1} s\{1/(2\Delta)\}^{R\sum_{j=1}^D p_j} \prod_{j=1}^{D}\prod_{r=1}^{R_0}\left[b_\lambda + \sum_{i_j=1}^{p_j}\{(\beta_{ji_j}^{0(r)})^2 + 2\Delta^2\}^{1/2}\right]^{a_\lambda+p_j} (RD)^{a_\lambda+1} \\ &\quad \times b_\lambda^{a_\lambda} b_\tau^{-a_\lambda}\prod_{j=1}^{D}\left(b_\lambda + 2^{1/2}p_j\Delta\right)^{(a_\lambda+p_j)(R-R_0)}\exp\{-n\epsilon_n^2 s^2(a_\lambda c/2 - 1)\}ds\{\Gamma(a_\tau + a_\lambda)/\Gamma(a_\tau)\} \\ &\leq \{5D/(2n^{1/2}\epsilon_n)\}^{R\sum_{j=1}^D p_j} C^{-1}(2\pi R)^{R\sum_{j=1}^D p_j/2} \prod_{j=1}^{D}\{\Gamma(a_\lambda)/\Gamma(a_\lambda + p_j)\}^R \\ &\quad \times \left[\int s\{1/(2\Delta)\}^{2R\sum_{j=1}^D p_j}\exp\{-n\epsilon_n^2 s^2(a_\lambda c/2 - 1)\}ds\right]^{1/2} b_\lambda^{a_\lambda} b_\tau^{-a_\lambda}(RD)^{a_\lambda+1} \\ &\quad \times \left[\int s\prod_{j=1}^{D}\prod_{r=1}^{R_0}\left[b_\lambda + \sum_{i_j=1}^{p_j}\{(\beta_{ji_j}^{0(r)})^2 + 2\Delta^2\}^{1/2}\right]^{2a_\lambda+2p_j}\prod_{j=1}^{D}\left(b_\lambda + 2^{1/2}p_j\Delta\right)^{2(a_\lambda+p_j)(R-R_0)}\right. \\ &\quad \left. \times \exp\{-n\epsilon_n^2 s^2(a_\lambda c/2 - 1)\}ds\right]^{1/2}\{\Gamma(a_\tau + a_\lambda)/\Gamma(a_\tau)\}. \end{aligned} \tag{11}$$

Using the Cauchy bound on Δ and Lagrange bound on $1/\Delta$, one obtains

$$\Delta \leq 1 + \epsilon_n s + D(M+1)^D, \quad 1/\Delta \leq 1 + D(M+1)^D + \epsilon_n s.$$

Next,

$$
\begin{aligned}
C_1 &= \int s \prod_{j=1}^{D} \prod_{r=1}^{R_0} \left[b_\lambda + \sum_{i_j=1}^{p_j} \{(\beta_{ji_j}^{0(r)})^2 + 2\Delta^2\}^{1/2} \right]^{2a_\lambda + 2p_j} \prod_{j=1}^{D} (b_\lambda + 2^{1/2} p_j \Delta)^{2(a_\lambda + p_j)(R - R_0)} \\
&\qquad \times \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} ds \\
&\leq \int_{s>1} s \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} \{b_\lambda + 2^{1/2} L(M + 1 + D(M+1)^D + \epsilon_n s)\}^{2(a_\lambda + L)RD} ds \\
&\leq \sum_{\ell=0}^{2RD(a_\lambda+L)} \binom{2RD(a_\lambda + L)}{\ell} (b_\lambda + 2^{1/2} L\tilde{M})^{2RD(a_\lambda+L)-\ell} (2^{1/2} L\epsilon_n)^\ell \frac{\Gamma(\ell/2 + 1)}{\{n\epsilon_n^2 (a_\lambda c/2 - 1)\}^{\ell/2+1}} \\
&= \Gamma\{2RD(a_\lambda + L)\} / \{n\epsilon_n^2 (a_\lambda c/2 - 1)\} [b_\lambda + 2^{1/2} L\tilde{M} + 2^{1/2} L / \{n(a_\lambda c/2 - 1)\}^{1/2}]^{2RD(a_\lambda+L)},
\end{aligned}
$$

where $\tilde{M} = M + 1 + D(M+1)^D$. Furthermore,

$$
\begin{aligned}
C_2 &= \int_{s>1} s \{1/(2\Delta)\}^{2R\sum_{j=1}^{D} p_j} \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} ds \\
&\leq \int_{D(M+1)^D \leq \epsilon_n s} s \{1/(2\Delta)\}^{2R\sum_{j=1}^{D} p_j} \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} ds \\
&\qquad + \int_{D(M+1)^D/\epsilon_n > s > 1} s \{1/(2\Delta)\}^{2R\sum_{j=1}^{D} p_j} \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} ds.
\end{aligned}
$$

Observe that $D(M+1)^D \leq \epsilon_n s$ implies $\Delta \geq 1$. Hence

$$
\int_{D(M+1)^D \leq \epsilon_n s} s\{1/(2\Delta)\}^{2R\sum_{j=1}^{D} p_j} \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} ds \leq 1/\{n\epsilon_n^2 (a_\lambda c/2 - 1)\}.
$$

On the other hand,

$$
\begin{aligned}
&\int_{D(M+1)^D/\epsilon_n > s > 1} s \{1/(2\Delta)\}^{2R\sum_{j=1}^{D} p_j} \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} ds \\
&\qquad \leq \int_{D(M+1)^D/\epsilon_n > s > 1} s\{1 + D(M+1)^D + \epsilon_n s\}^{2R\sum_{j=1}^{D} p_j} \exp\{-n\epsilon_n^2 s^2 (a_\lambda c/2 - 1)\} ds \\
&\qquad = \sum_{\ell=0}^{2R\sum_{j=1}^{D} p_j} \binom{2R\sum_{j=1}^{D} p_j}{\ell} \{1 + D(M+1)^D\}^{2R\sum_{j=1}^{D} p_j - \ell} (\epsilon_n)^\ell \frac{\Gamma(\ell/2 + 1)}{\{n\epsilon_n^2 (a_\lambda c/2 - 1)\}^{\ell/2+1}}
\end{aligned}
$$

$$
\begin{aligned}
&\times \int sC_s ds \leq \{5D/(2n^{1/2}\epsilon_n)\}^{R\sum_{j=1}^{D} p_j} C^{-1} (2\pi R)^{R\sum_{j=1}^{D} p_j/2} \prod_{j=1}^{D} \{\Gamma(a_\lambda)/\Gamma(a_\lambda + p_j)\}^R \\
&\times b_\tau^{-a_\lambda} b_\lambda^{a_\lambda} \frac{\Gamma\{2RD(a_\lambda + L)\}^{1/2}}{\{n\epsilon_n^2 (ca_\lambda/2 - 1)\}^{1/2}} \left[b_\lambda + 2^{1/2} L\tilde{M} + \frac{2^{1/2} L}{\{n(ca_\lambda/2 - 1)\}^{1/2}} \right]^{RD(a_\lambda+L)} \\
&\times \left\{ \frac{1}{n\epsilon_n^2 (ca_\lambda/2 - 1)} + \{1 + D(M+1)^D + n^{-1/2}\}^{2R\sum_{j=1}^{D} p_j} \frac{\Gamma(2R\sum_{j=1}^{D} p_j)}{\{n\epsilon_n^2 (ca_\lambda/2 - 1)\}} \right\}^{1/2} \\
&\times (RD)^{a_\lambda+1} \{\Gamma(a_\tau + a_\lambda)/\Gamma(a_\tau)\}.
\end{aligned}
$$

Using similar calculations and by invoking Lemma 2, we get

$$\begin{aligned}
\int sD_s ds \leq& \int s \exp\{n\epsilon_n^2 s^2 + \mathcal{E}(\epsilon_n s)\}\mathcal{D}(2n^{1/2}\epsilon_n s, \mathcal{F}_s, n^{1/2}\|\cdot\|_n)\exp(-2s^2 n\epsilon_n^2) \\
&\leq \{5D/(2n^{1/2}\epsilon_n)\}^{R\sum_{j=1}^D p_j} C^{-1}(2\pi R)^{R\sum_{j=1}^D p_j/2}\prod_{j=1}^D \{\Gamma(a_\lambda)/\Gamma(a_\lambda+p_j)\}^R (5D)^{RD\sum_{j=1}^D p_j} \\
&\quad \times\Bigg[\int s\{1/(2\Delta)\}^{2R\sum_{j=1}^D p_j}\exp\Big\{-n\epsilon_n^2 s^2\Big(1-cDR\sum_{j=1}^D p_j/2\Big)\Big\}ds\Bigg]^{1/2} \\
&\quad \times\Bigg[\int s\prod_{j=1}^D\prod_{r=1}^{R_0}\Big[b_\lambda+\sum_{i_j=1}^{p_j}\{(\beta_{ji_j}^{0(r)})^2+2\Delta^2\}^{1/2}\Big]^{2a_\lambda+2p_j}\prod_{j=1}^D(b_\lambda+2^{1/2}p_j\Delta)^{2(a_\lambda+p_j)(R-R_0)} \\
&\quad \times\exp\Big\{-n\epsilon_n^2 s^2\Big(1-cDR\sum_{j=1}^D p_j/2\Big)\Big\}ds\Bigg]^{1/2} \\
&\leq \{5D/(2n^{1/2}\epsilon_n)\}^{R\sum_{j=1}^D p_j} C^{-1}(2\pi R)^{R\sum_{j=1}^D p_j/2}\prod_{j=1}^D \{\Gamma(a_\lambda)/\Gamma(a_\lambda+p_j)\}^R (5D)^{RD\sum_{j=1}^D p_j} \\
&\quad \times\Big[\frac{\Gamma\{2RD(a_\lambda+L)\}}{\{n\epsilon_n^2(1-cDR\sum_{j=1}^D p_j/2)\}}\Big[b_\lambda+2^{1/2}L\tilde{M}+\frac{2^{1/2}L}{\{n(1-cDR\sum_{j=1}^D p_j/2)\}^{1/2}}\Big]^{2RD(a_\lambda+L)}\Big]^{1/2} \\
&\quad \times\Bigg[\frac{1}{n\epsilon_n^2(1-cRD\sum_{j=1}^D p_j/2)}+\{1+D(M+1)^D+n^{-1/2}\}^{2R\sum_{j=1}^D p_j} \\
&\quad \times\frac{\Gamma(2R\sum_{j=1}^D p_j)}{\Big\{n\epsilon_n^2(1-cDR\sum_{j=1}^D p_j/2)\Big\}}\Bigg]^{1/2}. \qquad (12)
\end{aligned}$$

Choosing $\epsilon_n = n^{-1/2}$, we get $\int sA_s ds \leq 1/2$, $\int sT_s ds \leq 8$. Furthermore,

$$\begin{aligned}
\int sC_s ds \leq& (5D/2)^{dR\sum_{j=1}^D p_j} C^{-1}(2\pi R)^{R\sum_{j=1}^D p_j/2}\prod_{j=1}^D\{\Gamma(a_\lambda)/\Gamma(a_\lambda+p_j)\}^R \\
&\times\Gamma(a_\tau+a_\lambda)b_\tau^{-a_\lambda}\Gamma(a_\tau)^{-1}\Gamma\{2RD(a_\lambda+L)\}^{1/2}\{(ca_\lambda/2-1)\}^{-1/2}(RD)^{a_\lambda+1}b_\lambda^{a_\lambda} \\
&\times[b_\lambda+2^{1/2}L\tilde{M}+2^{1/2}L\{n(ca_\lambda/2-1)\}]^{RD(a_\lambda+L)} \\
&\times\Big[(ca_\lambda/2-1)^{-1/2}+\{1+D(M+1)^D+n^{-1/2}\}^{2R\sum_{j=1}^D p_j} \\
&\quad\times\Gamma\Big(2R\sum_{j=1}^D p_j\Big)/\{ca_\lambda/2-1)\}\Big]^{1/2} \\
&= C_n
\end{aligned}$$

and

$$
\begin{aligned}
\int sD_s ds \leq{} & (5D/2)^{dR\sum_{j=1}^{D} p_j} C^{-1} (2\pi R)^{R\sum_{j=1}^{D} p_j/2} \prod_{j=1}^{D} \{\Gamma(a_\lambda)/\Gamma(a_\lambda + p_j)\}^R \\
& \times \Gamma\{2RD(a_\lambda + L)\}^{1/2} \Big\{\Big(1 - cDR \sum_{j=1}^{D} p_j/2\Big)\Big\}^{-1/2} (5D)^{RD\sum_{j=1}^{D} p_j} \\
& \times \left[b_\lambda + 2^{1/2} L\tilde{M} + 2^{1/2} L \Big\{ n\Big(1 - cDR \sum_{j=1}^{D} p_j/2\Big)\Big\}^{-1/2} \right]^{RD(a_\lambda + L)} \\
& \times \Big[\Big(1 - cRD \sum_{j=1}^{D} p_j/2\Big)^{-1} + \{1 + D(M+1)^D + n^{-1/2}\}^{2R\sum_{j=1}^{D} p_j} \\
& \Gamma\Big(2R \sum_{j=1}^{D} p_j\Big) \Big/ \Big\{1 - cDR \sum_{j=1}^{D} p_j/2\Big\}\Big]^{1/2} \\
={} & D_n.
\end{aligned}
$$

This concludes the proof of Theorem 2. □

References

[1] A. Armagan, D.B. Dunson, J. Lee, Generalized double Pareto shrinkage, Statistica Sinica 23 (2013) 119–143.

[2] A. Armagan, D.B. Dunson, J. Lee, W.U. Bajwa, N. Strawn, Posterior consistency in linear models under shrinkage priors, Biometrika 100 (2013) 1011–1018.

[3] C.M. Carvalho, N.G. Polson, J.G. Scott, The horseshoe estimator for sparse signals, Biometrika 97 (2010) 465–480.

[4] W. Chu, Z. Ghahramani, Probabilistic models for incomplete multi-dimensional arrays, AISTATS 2009, pp. 89–96.

[5] C.M. Crainiceanu, A.M. Staicu, C.Z. Di, Generalized multilevel functional regression, J. Amer. Statist. Assoc. 104 (2009) 1550–1561.

[6] S. Gandy, B. Recht, I. Yamada, Tensor completion and low-n-rank tensor recovery via convex optimization, Inverse Problems 27 (2011) 025010.

[7] R. Guhaniyogi, S. Qamar, D.B. Dunson, Bayesian tensor regression, arXiv preprint arXiv:1509.06490.

[8] P.D. Hoff, Multilinear tensor regression for longitudinal relational data, Ann. Appl. Statist. 9 (2015) 1169–1193.

[9] T.G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (2009) 455–500.

[10] B. Laurent, P. Massart, Adaptive estimation of a quadratic functional by model selection, Ann. Statist. 28 (2000) 1302–1338.

[11] N. Lazar, The Statistical Analysis of Functional MRI Data, Springer Science & Business Media, 2008.

[12] J. Liu, P. Musialski, P. Wonka, J. Ye, Tensor completion for estimating missing values in visual data, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2013) 208–220.

[13] S. Negahban, M.J. Wainwright, Restricted strong convexity and weighted matrix completion: Optimal bounds with noise, J. Machine Learning Res. 13 (May) (2012) 1665–1697.

[14] N.G. Polson, J.G. Scott, Local shrinkage rules, lévy processes and regularized regression, J. Roy. Statist. Soc. Ser. B 74 (2012) 287–311.

[15] P.T. Reiss, L. Huo, Y. Zhao, C. Kelly, R.T. Ogden, Wavelet-domain regression and predictive inference in psychiatric neuroimaging, Ann. Appl. Statist. 9 (2015) 1076–1101.

[16] P.T. Reiss, R.T. Ogden, Functional generalized linear models with images as predictors, Biometrics 66 (2010) 61–69.

[17] T. Suzuki, Convergence rate of Bayesian tensor estimatior and its minimax optimality, Proceedings of the 32nd International Conference on Machine Learning (2015) 1273–1282.

[18] R. Tomioka, T. Suzuki, Convex tensor decomposition via structured Schatten norm regularization, Advances in Neural Information Processing Systems (2013) 1331–1339.

[19] A.W. van der Vaart, H. Van Zanten, Information rates of nonparametric Gaussian process methods, J. Machine Learning Res. 12 (Jun) (2011) 2095–2119.

[20] J. Zhou, A. Bhattacharya, A.H. Herring, D.B. Dunson, Bayesian factorizations of big sparse tensors, J. Amer. Statist. Assoc. 110 (2015) 1562–1576.

[21] H. Zhou, L. Li, H. Zhu, Tensor regression with applications in neuroimaging data analysis, J. Amer. Statist. Assoc. 108 (2013) 540–552.