ORIGINAL ARTICLE

Convergence Rate for Predictive Densities of Bayesian Generalized Linear Models with a Scalar Response and Symmetric Tensor Predictor

Rajarshi Guhaniyogi[1] | **Sharmistha Guha**[2]

[1]Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, 95064, USA

[2]Department of Statistical Science, 206 Old Chemistry Building, Duke University, Durham, NC, 27708, USA

Correspondence
Rajarshi Guhaniyogi, Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, 95064, USA
Email: rguhaniy@ucsc.edu

Funding information

This article investigates statistical convergence rates for predictive densities of a novel Bayesian generalized linear model (GLM) framework with a scalar response and a symmetric tensor predictor with labeled “nodes." Such a framework may appear in a variety of applications, including diffusion weighted magnetic resonance imaging (DWI) and functional magnetic resonance imaging (fMRI), among others. This article specifically focuses on a class of models where the over-arching goal is to identify influential nodes and cells of the symmetric tensor. We establish a near optimal convergence rate for the posterior predictive density from the proposed model to the true density, depending on how the number of tensor nodes grows with the sample size. Moreover, we show that the method has adaptivity to the unknown rank of the true tensor, i.e., the near optimal rate is achieved even if the rank of the true tensor coefficient is not known a priori.

KEYWORDS
Low-rank tensor decomposition; Posterior convergence rate; Symmetric tensor; Spike-and-slab prior

1 | INTRODUCTION

Of late, scientific applications often involve predictors having a multidimensional array or tensor structure, which are higher order analogues to vectors and matrices. Analogous to the rows and columns of a matrix, various axes of a tensor are known as tensor modes and the indices of a tensor mode are often referred to as “tensor nodes." Entries in a tensor are known as “tensor cells." This article considers *symmetric* tensors, which are invariant upon interchanging the modes. We specifically focus on developing a regression relationship between a scalar response and a symmetric tensor predictor, with the ability of identifying tensor nodes influential in predicting the response. One major application of such modeling framework appears in brain connectome data, where the goal is to predict a brain related phenotype from the brain connectome network of subjects, with an emphasis of drawing inference on brain regions of interests(ROIs) related to the phenotype (Guha and Rodriguez, 2018; Relión et al., 2019).

In developing a modeling approach to address our problem of interest, one can possibly proceed to vectorize the symmetric tensor response and regress it on the predictors, leading to a high dimensional vector regression problem (Craddock et al., 2009; Richiardi et al., 2011). This approach is able to make use of the expanding literature on Bayesian high dimensional regression (Park and Casella, 2008; Carvalho et al., 2010) but appears to be less than adequate to achieve all of our inferential goals simultaneously for a few reasons. First, the ordinary high dimensional regression framework assumes the coefficients corresponding to the tensor cells to be exchangeable, although, intuitively, the coefficients related to the same tensor node should be correlated a priori. Second, the strategy of reshaping a symmetric tensor into a vector leads to a massive dimensional vector predictor with applications often involving a limited number of samples. From an inferential point of view, Bayesian high-dimensional regression frameworks may be statistically inefficient when the number of predictors far exceeds the sample size (Armagan et al., 2013). More importantly, identification of important tensor nodes is not one of the inferential objectives of these approaches. Recent developments on tensor regression (Zhou et al., 2013; Guhaniyogi et al., 2017) provide a solution to the problem by exploiting the tensor structure of the predictor in the model and prior development. However, these approaches do not generally take into account the symmetry constraint in the tensor predictor, tend to focus mainly on prediction, and are not specifically designed to detect important nodes impacting the response.

A recent approach to address all the inferential objectives mentioned above is outlined in Guha and Rodriguez, 2018 in the context of symmetric predictor matrices. More specifically, Guha and Rodriguez, 2018 develop a novel shrinkage prior on the symmetric matrix coefficients by combining ideas from low-rank matrix factorization and the Bayesian shrinkage prior literature. The structure offers parsimony by allowing identification of important tensor node specific coefficient vectors using a spike-and-slab prior on them. The framework exhibits good empirical performance with precise predictive inference as well as accurate identification of important tensor nodes. Moreover, the proposed prior allows auto tuning of all the hyperparameters with Markov chain Monte Carlo chains showing reasonably rapid mixing. While Guha and Rodriguez, 2018 provide the methodological and empirical motivations regarding the prior construction, rigorous theoretical understanding of Bayesian symmetric tensor regressions is yet to be established. Furthermore, the modeling framework is introduced and tested with normally distributed response variables.

Motivated by Guha and Rodriguez, 2018, we propose a generalized linear modeling framework with a scalar response and a symmetric tensor predictor. We adopt a low-rank structure for the symmetric tensor predictor coefficient and assign a spike-and-slab prior on node specific latent vectors within the low-rank structure to determine the tensor nodes significantly related to the scalar response a posteriori. A major contribution of this article is in developing conditions on the ranks and magnitudes of the true tensor coefficients and the number of tensor nodes for the near optimal learning of the proposed GLM. Note that several influential articles have emerged in the last few years detailing conditions for posterior contraction in ordinary high dimensional regression models, both with various point-mass priors in the many normal-means models (Castillo et al., 2012; Belitser and Nurushev, 2015; Martin et al., 2017), and with classes of continuous shrinkage priors (Song and Liang, 2017; Wei and Ghosal, 2017). In contrast, there is a dearth of papers studying posterior contraction properties for generalized linear models with tensor predictors in the Bayesian paradigm. A few recent articles (Guhaniyogi, 2017; Guhaniyogi et al., 2017) offer conditions for consistency or optimal rates for posterior contraction with tensor predictors without the symmetry constraint, and with a different class of multiway shrinkage priors (Guhaniyogi et al., 2017). As a result, the theoretical construction in Guhaniyogi et al., 2017 does not find ready extension to our framework. Additionally, we relax the key assumption in Guhaniyogi et al., 2017 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.

The current article is based upon a more realistic assumption that the rank of the fitted tensor coefficient is greater than or equal to the rank of the true tensor coefficient.

The rest of the article proceeds as follows. Section 2 develops the notations, defines the GLM framework for the fitted model and the true data generating model, and details the prior distributions on the parameters. Section 3 describes the posterior contraction rate results for the predictive distribution. Finally Section 4 concludes the article with an eye towards future work.

2 | PROBLEM SETTING

2.1 | Notations

A $D-$way tensor $\boldsymbol{\Gamma} \in \otimes_{l=1}^{D} \mathbb{R}^{V_l}$ is a multidimensional array whose $(k_1, ..., k_D)$th *cell* is denoted by $\Gamma_{(k_1,...,k_D)}$, $1 \leq k_1 \leq V_1$,...,$1 \leq k_D \leq V_D$. When $D = 2$, a tensor corresponds to a matrix. This article mainly focuses on symmetric tensors with dummy diagonal entries (set at 0 for definiteness) ensuring $V_1 = \cdots = V_D = V$ and $\Gamma_{(k_1,...,k_D)} = \Gamma_{(P(k_1),...,P(k_D))}$, for any permutation $P(\cdot)$ of $\{k_1, ..., k_D\}$ and $\Gamma_{(k_1,...,k_D)} = 0$, if any two of the indices k_l and $k_{l'}$ are equal. Similar to row and column indices of a matrix, the indices $\mathcal{N} = \{1, 2, ..., V\}$ for symmetric tensors are referred to as *tensor nodes*. Let $\mathcal{K} = \{(k_1, ..., k_D) : 1 \leq k_1 < \cdots < k_D \leq V\}$ be a set of indices with cardinality $q = \frac{V(V-1)\cdots(V-D+1)}{D!}$. While expressing a symmetric tensor $\boldsymbol{\Gamma}$ with 0 diagonal entries, it is enough to specify $\Gamma_{\boldsymbol{k}}$ for $\boldsymbol{k} \in \mathcal{K}$. This holds since for any $\boldsymbol{k} \notin \mathcal{K}$ $\exists$ a permutation $P(\cdot)$ s.t. $(P(k_1), ..., P(k_D)) \in \mathcal{K}$. Then, by the property of the symmetric tensor, $\Gamma_{P(\boldsymbol{k})} = \Gamma_{\boldsymbol{k}}$. A symmetric tensor with 0 diagonal entries $\boldsymbol{\Gamma}$ assumes a rank-1 PARAFAC decomposition if $\boldsymbol{\Gamma}_{\boldsymbol{k}}$ for $\boldsymbol{k} \in \mathcal{K}$ can be expressed as $\Gamma_{\boldsymbol{k}} = \gamma_{k_1} \cdots \gamma_{k_D}$, for $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_V)' \in \mathbb{R}^V$. A rank R symmetric PARAFAC decomposition expresses $\Gamma_{\boldsymbol{k}}$ as $\Gamma_{\boldsymbol{k}} = \sum_{r=1}^{R} \gamma_{k_1}^{(r)} \cdots \gamma_{k_D}^{(r)}$, where $\boldsymbol{\gamma}^{(r)} = (\gamma_1^{(r)}, ..., \gamma_V^{(r)})' \in \mathbb{R}^V$. Importantly, for two symmetric tensors $\boldsymbol{A}$ and $\boldsymbol{B}$ with zero diagonal entries, the Frobenius inner product between $\boldsymbol{A}$ and $\boldsymbol{B}$ are given by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = D! \sum_{\boldsymbol{k} \in \mathcal{K}} a_{\boldsymbol{k}} b_{\boldsymbol{k}}$. Finally, $|| \cdot ||$ and $|| \cdot ||_\infty$ denote the l_2 and l_∞ norms, respectively, for both vectors and higher order tensors.

2.2 | Modeling Framework

For $i = 1, ..., n$, let y_i be the scalar response and $\boldsymbol{X}_i = ((x_{i,(k_1,...,k_D)}))_{k_1,...,k_D=1}^{V} \in \mathbb{R}^{V \times \cdots \times V}$ denote the symmetric tensor response with 0 diagonal entries. We assume that the data are generated from the generalized linear model given by the following density function

$$g_0(y_i | \boldsymbol{X}_i) = \exp(a(\alpha_0) y_i + b(\alpha_0) + c(y_i)), \ \alpha_0 = \sum_{\boldsymbol{k} \in \mathcal{K}} x_{i,\boldsymbol{k}} \Gamma_{0,\boldsymbol{k}}, \tag{1}$$

where $\Gamma_{0,\boldsymbol{k}}$ corresponds to the $\boldsymbol{k} = (k_1, ..., k_D)$-th entry of a symmetric tensor (with 0 diagonal entries) $\boldsymbol{\Gamma}_0$, $a(h)$ and $b(h)$ are continuously differentiable functions, with $a(h)$ having a nonzero derivative. This parameterization includes some popular classes of densities, including binary logit and probit regressions of y on $\boldsymbol{X}$, Poisson regression of y on $\boldsymbol{X}$ with count valued response, and normal regression with known error variance for continuous response y (Jiang, 2007). The conditional density of y_i given $\boldsymbol{X}_i$ fitted to the data is also assumed to belong to the same class of generalized linear models, and is given by

$$g(y_i | \boldsymbol{X}_i) = \exp(a(\alpha) y_i + b(\alpha) + c(y_i)), \ \alpha = \sum_{\boldsymbol{k} \in \mathcal{K}} x_{i,\boldsymbol{k}} \Gamma_{\boldsymbol{k}}, \tag{2}$$

where $\Gamma_{\boldsymbol{k}}$ is the $\boldsymbol{k}$-th entry of $\boldsymbol{\Gamma}$, which is a symmetric tensor with 0 diagonal entries.

Suppose $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}$ assume symmetric rank-R_0 and rank-R PARAFAC decompositions, respectively, for $R > R_0$, so that

$$\Gamma_{0,\boldsymbol{k}} = \sum_{r=1}^{R_0} \gamma_{0,k_1}^{(r)} \cdots \gamma_{0,k_D}^{(r)},\ \Gamma_{\boldsymbol{k}} = \sum_{r=1}^{R} \lambda_r \gamma_{k_1}^{(r)} \cdots \gamma_{k_D}^{(r)}, \tag{3}$$

where $\boldsymbol{\gamma}^{(r)} = (\gamma_1^{(r)}, ..., \gamma_V^{(r)})$ and $\boldsymbol{\gamma}_0^{(r)} = (\gamma_{0,1}^{(r)}, ..., \gamma_{0,V}^{(r)}) \in \mathbb{R}^V$ for all $r = 1, ..., R$. Since the rank of the fitted symmetric tensor coefficient $\boldsymbol{\Gamma}$ is assumed to be higher than the rank of the true tensor coefficient $\boldsymbol{\Gamma}_0$, rank specific binary inclusion variables $\lambda_r \in \{0, 1\}$ are added in order to *switch-off* the contribution of unnecessary summands. The assumed low-rank decomposition offers parsimony by reducing the number of estimable parameters from $V(V-1)\cdots(V-D+1)/D!$ to RV, typically with $R \ll V$. When $D = 2$, the formulation assumes further simplification. To see this, denote $\tilde{\boldsymbol{\gamma}}_h = (\gamma_h^{(1)}, ..., \gamma_h^{(R)})'$, $h = 1, ..., V$ and $\boldsymbol{\Lambda} = diag(\lambda_1, .., \lambda_R)$. The $\boldsymbol{k} = (k_1, k_2)$th entry of $\boldsymbol{\Gamma}$ then simplifies as $\Gamma_{\boldsymbol{k}} = \tilde{\boldsymbol{\gamma}}_{k_1}' \boldsymbol{\Lambda} \tilde{\boldsymbol{\gamma}}_{k_2}$, $\boldsymbol{k} \in \mathcal{K}$, which represents a *bilinear* (Hoff, 2005) interaction between $\tilde{\boldsymbol{\gamma}}_{k_1}$ and $\tilde{\boldsymbol{\gamma}}_{k_2}$. Accordingly, the significance of the $\boldsymbol{k}$th tensor cell of $\boldsymbol{\Gamma}$ in explaining the response increases with the similarity in the positions of $\tilde{\boldsymbol{\gamma}}_{k_1}$ and $\tilde{\boldsymbol{\gamma}}_{k_2}$, the similarity being measured by the weighted dot product between these two variables in the latent space.

From (3), the hth tensor node of the symmetric tensor predictor $\boldsymbol{X}$ is deemed to have no impact on the response if $\tilde{\boldsymbol{\gamma}}_h = \boldsymbol{0}$, $h \in \mathcal{N}$. The $\boldsymbol{k}$th cell is considered unrelated to the response if $\Gamma_{\boldsymbol{k}} = 0$. Since $\Gamma_{\boldsymbol{k}} = 0$ if $\tilde{\boldsymbol{\gamma}}_{k_l} = 0$ for some k_l, the proposed formulation assumes that the contribution of the $\boldsymbol{k}$th cell of the tensor predictor to the response is insignificant if k_lth node is unrelated to the response, for some k_l.

Example 1 (Brain Connectome Data): In many neuroscientific applications, it is of interest to build a predictive model of a brain related phenotype (e.g., presence of a neuronal disease) on the connectivity network in a human brain (referred to as the *brain connectome*) (for e.g., see Relión et al., 2019). 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 a brain imaging technique known as *diffusion tensor imaging* (DTI). Alternatively, the brain connectome tensor can also be constructed by computing the correlation of *functional magnetic resonance imaging* (fMRI) signals for different pairs of regions after suitably thresholding them to zero below a certain pre-specified cut-off. The inferential interest here lies in predicting the phenotypic response from the brain connectome matrix, as well as identifying ROIs significantly related to the response. This appears to be a direct application of (1), with y and $\boldsymbol{X}$ as the phenotype and the symmetric brain connectome matrix, respectively, and nodes in the matrix representing the ROIs.

Example 2 (International Trade Data): Developing a regression relationship between world *gross domestic product* (GDP) and multilateral trade between countries is an informative exercise in international trade theory. Analysis of datasets with such information is important to statistically identify countries which are major economic drivers of the world, and also to direct significant world economic policies by international financial institutions (Chan and Kuo, 2005; Hufbauer, 2005). In the context of (1), the response and predictors would be the world GDP and multilateral trade relationships (which constitute a symmetric higher order tensor), respectively. The countries are the tensor nodes to draw inference on. In this context, it is generally believed that free trade agreements between countries could benefit the overall economic health of the world. For example, one can consider the trilateral free trade agreement between China-Japan-South Korea (Chiang, 2013), or between the U.S.-Canada-Mexico (referred to as the North Atlantic Free Trade Agreement or NAFTA) (Brown et al., 1995). It is instructive to statistically analyze important economic outcomes like GDP in relation to such multi-lateral free trade agreements.

2.3 | Prior Structure

To assess if the hth tensor node is active in predicting the response, we assign a spike-and-slab mixture prior distribution on $\tilde{\gamma}_h$ as

$$\tilde{\gamma}_h \sim \zeta_h N(\mathbf{0}, \boldsymbol{I}) + (1-\zeta_h)\delta_{\mathbf{0}},\ \ \zeta_h \sim Ber(\Delta),\ \Delta \sim U(0,1), \tag{4}$$

where $\delta_{\mathbf{0}}$ is the Dirac-delta function at $\mathbf{0}$, Δ corresponds to the probability of the nonzero mixture component and ζ_h is a binary indicator set to 0 if $\tilde{\gamma}_h = \mathbf{0}$. Thus, the posterior distributions of the ζ_h's are analyzed to ascertain which nodes are influential in predicting the response. Finally, to infer on how many ranks are necessary to express $\boldsymbol{\Gamma}$, the rank specific binary inclusion variables, the λ_r's, are assigned a hierarchical prior, $\lambda_r \sim Ber(\nu_r)$, $\nu_r \sim Beta(1, r^{\eta})$. Choosing $\eta > 1$ ensures increasing shrinkage on λ_r as r grows. Thus a low-rank solution to $\boldsymbol{\Gamma}$ is favored a priori, which helps avoid over-fitting.

Analysis of datasets using the model (2) involving a continuous scalar response and symmetric tensor predictors are available in some recent work Guha and Rodriguez (2018), though a rigorous theoretical treatment of such models is missing in the literature. The overarching goal of this article is to develop theoretical conditions to draw optimal predictive inference from such models. It will be shown in due course that the posterior predictive loss (defined in Section 3) of our model decays at the “near" optimal rate to 0 under fairly mild assumptions. Moreover, such theoretical results will be obtained for an easily computable posterior with standard Markov chain Monte Carlo updates for all the parameters.

3 | CONVERGENCE RATE ANALYSIS

This article assesses the predictive accuracy of the proposed model $g(y|\boldsymbol{X})$ in estimating the true model $g_0(y|\boldsymbol{X})$, following the notion of convergence described in Jiang, 2007. Define the Hellinger distance between g and g_0 as

$$d_H(g, g_0) = \sqrt{\int\int (\sqrt{g(y|\boldsymbol{X})} - \sqrt{g_0(y|\boldsymbol{X})})^2 \nu_y(dy)\nu_{\boldsymbol{X}}(d\boldsymbol{X})},$$

where $\nu_{\boldsymbol{X}}$ is the unknown probability measure for $\boldsymbol{X}$, and ν_y is the dominating measure for g and g_0. We focus on showing $E_{g_0}\Pi[d_H(g, g_0) > \epsilon_n | \{y_i, \boldsymbol{X}_i\}_{i=1}^n] < \xi_n$, for large n, for some sequences ϵ_n, ξ_n converging to 0 as $n \to \infty$, where $\Pi(\mathcal{S}|\{y_i, \boldsymbol{X}_i\}_{i=1}^n)$ is the posterior probability of the set $\mathcal{S}$. The result implies that the posterior probability outside of a shrinking neighborhood around the true predictive density g_0 converges to 0 as $n \to \infty$. Specifically, we focus on identifying conditions that lead to convergence rate ϵ_n of the order of $n^{-1/2}$ upto a $\log(n)$ factor.

3.1 | Framework and Main Results

Without loss of generality, the predictor $\boldsymbol{X}_i$ satisfies $|x_{i,\boldsymbol{k}}| < 1$ for all i and $\boldsymbol{k} \in \mathcal{K}$. In what follows, we add the subscript n to the number of tensor nodes V_n, the rank R_n of $\boldsymbol{\Gamma}$ and rank $R_{0,n}$ of the true symmetric tensor coefficient $\boldsymbol{\Gamma}_0$. We assume V_n, R_n and $R_{0,n}$ are all non-decreasing functions of n, with $R_n < V_n$ and $R_n > R_{0,n}$ for all large n. Hence, the number of elements in $\mathcal{K}$, given by $q_n = V_n(V_n-1)...(V_n-D+1)/D!$, is a function of n. This paradigm attempts to capture the fact that q_n grows much faster than n, and a higher rank CP decomposition of $\boldsymbol{\Gamma}$ can be estimated more precisely in the presence of a larger sample size n.

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 $\mathcal{E}_n(\kappa)$, defined, for each $\kappa > 0$ by

$$\mathcal{E}_n(\kappa) = -\log\left\{\Pi(||\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_0||_\infty \leq \kappa)\right\}. \tag{5}$$

In order to achieve an optimal rate of convergence for the posterior, one expects the prior to put considerable mass around $\mathbf{\Gamma}_0$. Since $\mathbf{\Gamma}_0$ is not known, it is not desirable to have a lot of prior mass around one point or a few points. Rather, the prior mass should be spread judiciously, taking into account the wide range of possibilities for $\mathbf{\Gamma}_0$. Prior concentration provides such a quantification of prior mass around the truth. Instead of characterizing the prior concentration function $\mathcal{E}_n(\kappa)$, we evaluate the prior concentration conditional on a set C given by $C = \left\{\lambda_1 = 1, ..., \lambda_{R_{0,n}} = 1, \lambda_{R_{0,n}+1} = 0, ..., \lambda_{R_n} = 0\right\}$, with Lemma 1 quantifying a lower bound on $P(C)$.

Lemma 1 *For $\lambda_r|v_r \overset{ind.}{\sim} Ber(v_r)$ and $v_r \sim Beta\,(1, r^\eta)$, $r = 1, ..., R_n$,*

$$P(C) = P(\lambda_1 = 1, ..., \lambda_{R_{0,n}} = 1, \lambda_{R_{0,n}+1} = 1, ..., \lambda_{R_n} = 1) \geq \frac{R_{0,n}^{\eta(R_n - R_{0,n})}}{(1 + R_{0,n}^\eta)^{R_n}}.$$

The prior concentration conditional on the set C is given by

$$\mathcal{E}_n(\kappa|C) = -\log\left\{\Pi(||\mathbf{\Gamma} - \mathbf{\Gamma}_0||_\infty \leq \kappa|C)\right\} \tag{6}$$

Lemma 2 below presents an upper bound on the conditional prior concentration corresponding to our proposed prior distribution in Section 2.3.

Lemma 2 *Let $\tilde{\gamma}_{0,h} = (\gamma_{0,h}^{(1)}, ..., \gamma_{0,h}^{(R_{0,n})})'$ and for $\boldsymbol{k} \in \mathcal{K}$, let $u_{\boldsymbol{k},n}$ be the only positive root of the equation*

$$x\prod_{s=2}^{D}(x + ||\tilde{\gamma}_{0,k_s}||) + ||\tilde{\gamma}_{0,k_1}||x\prod_{s=3}^{D}(x + ||\tilde{\gamma}_{0,k_s}||) + \cdots + x\prod_{s=1}^{D-1}||\tilde{\gamma}_{0,k_s}|| = \upsilon_n. \tag{7}$$

Assume $u_n = min_{\boldsymbol{k}\in\mathcal{K}}\, u_{\boldsymbol{k},n}$. Then,

$$\begin{aligned}\mathcal{E}_n(\upsilon_n|C) \leq & \sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||^2/2 + \frac{R_{0,n}V_n}{2}\log(2\pi) + \log\left(\frac{R_{0,n}V_n + 1}{R_{0,n}V_n}\right) + R_{0,n}V_n\log\left(2R_{0,n}/u_n\right) \\ & + V_n u_n^2/R_{0,n}\end{aligned}$$

Define the function $H(\kappa) = 1 + \kappa \sup\limits_{|w|\leq\kappa} |a'(w)| \sup\limits_{|w|\leq\kappa} |b'(w)/a'(w)|$, where $a'(w)$ and $b'(w)$ are derivatives of the functions $a(w)$ and $b(w)$, respectively. We now state the main theorem involving the contraction of the fitted predictive density to the true predictive density.

Theorem 3 *For a sequence ϵ_n satisfying $0 < \epsilon_n < 1$, $n\epsilon_n^2 \to \infty$, and another sequence C_n, let the following conditions hold*

(a) $R_nV_n\log(V_n) = o(n\epsilon_n^2)$

(b) $R_nV_n\log(1/\epsilon_n^2) = o(n\epsilon_n^2)$

(c) $R_nV_n\log(H(R_nC_n^D)) = o(n\epsilon_n^2)$,

(d) $(1 - \Phi(C_n)) \leq e^{-4n\epsilon_n^2}$*, for all large n*

(e) $\lim_{n\to\infty}\sum_{h=1}^{V_n}||\tilde{\gamma}_{0,h}|| \leq \infty$*, where* $\tilde{\gamma}_{0,h} = (\gamma_{0,h}^{(1)}, ..., \gamma_{0,h}^{(R_n)})'$.

Then, $E_{g_0}\Pi\{d_H(g, g_0) > 4\epsilon_n|\{Y_i, x_i\}_{i=1}^n\} < 4e^{-n\epsilon_n^2}$, for all large n.

The following remarks characterize $H(\kappa)$ and its implications for various regression settings under GLM.

Remark 1: For ordinary linear regression with normal errors, $H(\kappa)$ grows at most at the order of $|\kappa|^2$. Thus, assumption (c) becomes equivalent to $R_n V_n \log(C_n) = o(n\epsilon_n^2)$, considering assumption (a).

Remark 2: For binary regression with logit or probit links, $H(\kappa)$ grows at most linearly with $|\kappa|$. Thus, assumption (c) becomes equivalent to $R_n V_n \log(C_n) = o(n\epsilon_n^2)$, considering assumption (a). For our theoretical exposition, we will focus on continuous and binary regression only. Theorem 3, together with the functional properties of $H(\kappa)$ mentioned, leads to the following result on the convergence rate ϵ_n of the proposed model.

Corollary 4 *Let,* $\lim_{n\to\infty}\sum_{h=1}^{V_n}||\tilde{\gamma}_{0,h}|| \leq \infty$, *where* $\tilde{\gamma}_{0,h} = (\gamma_{0,h}^{(1)}, ..., \gamma_{0,h}^{(R_n)})'$. *Assume that for some* $0 < \xi < 1$, $V_n \leq M_1 n^{\xi}$ *(for some constant* $M_1 > 0$*) and the tensor rank* R_n *grows at a much slower rate of* $(\log n)^{z_1}$ *for some* z_1*, i.e.,* $R_n \leq M_2(\log n)^{z_1}$*, for some constant* M_2*. Choose* C_n *such that* $n^{\phi_1} \leq C_n \leq n^{\phi_2}$*, satisfying* $0 < \xi/2 < \phi_1 < \phi_2$*. Then the convergence rate* ϵ_n *can be expressed as* $\epsilon_n \sim n^{-(1-\xi)/2}(\log n)^{z_1/2+1}$ *for the linear regression model, as well as the binary regression model with logistic or probit link functions.*

Remark 3: Note that the condition $\lim_{n\to\infty}\sum_{h=1}^{V_n}||\tilde{\gamma}_{0,h}|| \leq \infty$ includes as a special case the scenario in which only a fixed and finite number of $||\tilde{\gamma}_{0,h}||$'s are nonzero, while also allowing a more realistic setup with many small $||\tilde{\gamma}_{0,h}||$, none of which are exactly zero. The convergence rate also depends on how V_n and R_n grows with n. In fact, the convergence rate deteriorates as ξ becomes higher, i.e., the number of tensor nodes grows.

4 | CONCLUSION

This article investigates the convergence rate of the predictive distribution for generalized linear models involving a scalar response and a symmetric tensor predictor. Under mild assumptions, we provide a "near optimal" convergence rate for the predictive distribution of the proposed model. The theoretical results proved here allow the number of tensor cells to grow much faster than the sample size. The near optimal rate is rank adaptive, i.e., it holds even if the rank of the symmetric tensor coefficient for the true data generating regression model is unknown. Most importantly, the bound on the predictive accuracy is achieved for a prior that leads to an easily computable posterior.

Several future directions of research emerge from this article. For example, it might be of interest to relax assumption (e) in Theorem 3 and investigate convergence rate by allowing $\sum_{h=1}^{V_n}||\tilde{\gamma}_{0,h}||$ to vary slowly as an increasing function of n. Another interesting future direction constitutes extending this theoretical set up to prove the tensor node selection consistency for the proposed model.

APPENDIX A

We begin by stating a Lemma.

Lemma 5 *Let* $J(x)$ *be a monic polynomial given by* $J(x) = x^D + b_{D-1}x^{D-1} + \cdots + b_1 x - b_0$, $b_0, ..., b_{D-1} \geq 0$. *If* x_0 *is a real positive root of the equation* $J(x) = 0$, *then* $1/x_0 \leq 1 + (b_1/b_0)$.

Proof Let $z = 1/x$. Then $J(x) = 0$ implies $J(1/z) = 0$, i.e., $z^D - (b_1/b_0)z^{D-1} - \cdots - (b_{D-1}/b_0)z - (1/b_0) = 0$. Since this is a monic polynomial with $1/x_0$ as one of its positive real roots, by the Lagrange-Maclaurin theorem, $1/x_0 \leq 1 + (b_1/b_0)$.

Proof of Lemma 1

$P(\lambda_r = 1) = E(\nu_r) = \frac{1}{1+r^\eta}$ for $r = 1, ..., R_n$. Then,

$$P(\lambda_1 = 1, ..., \lambda_{R_{0,n}} = 1, \lambda_{R_{0,n}+1} = 0, ..., \lambda_{R_n} = 0) = \prod_{r=1}^{R_{0,n}} \frac{1}{(1+r^\eta)} \prod_{r=R_{0,n}+1}^{R_n} \frac{r^\eta}{(1+r^\eta)}$$
$$\geq \frac{1}{(1+R_{0,n}^\eta)^{R_{0,n}}} \left\{ \frac{R_{0,n}^\eta}{(1+R_{0,n}^\eta)} \right\}^{R_n - R_{0,n}} = \frac{R_{0,n}^{\eta(R_n - R_{0,n})}}{(1+R_{0,n}^\eta)^{R_n}}.$$

The first inequality follows due to the fact that $r^\eta/(1+r^\eta)$ is a monotone increasing function of r and $1/(1+r^\eta)$ is a monotone decreasing function of r.

Proof of Lemma 2

Let $\mathcal{J} = \left\{\mathbf{\Gamma} : ||\mathbf{\Gamma} - \mathbf{\Gamma}_0||_\infty \leq \upsilon_n\right\}$. Under $\mathcal{C}$, for $\boldsymbol{k} \in \mathcal{K}$,
$|\Gamma_{\boldsymbol{k}} - \Gamma_{0,\boldsymbol{k}}| = |\sum_{r=1}^{R_{0,n}} \gamma_{k_1}^{(r)} \cdots \gamma_{k_D}^{(r)} - \sum_{r=1}^{R_{0,n}} \gamma_{0,k_1}^{(r)} \cdots \gamma_{0,k_D}^{(r)}| = |\sum_{r=1}^{R_{0,n}} (\gamma_{k_1}^{(r)} - \gamma_{0,k_1}^{(r)}) \prod_{s=2}^{D} \gamma_{k_s}^{(r)}| + \cdots + |\sum_{r=1}^{R_{0,n}} (\gamma_{k_D}^{(r)} - \gamma_{0,k_D}^{(r)}) \prod_{s=1}^{D-1} \gamma_{0,k_s}^{(r)}| \leq$ $||\tilde{\gamma}_{k_1} - \tilde{\gamma}_{0,k_1}|| \prod_{s=2}^{D} ||\tilde{\gamma}_{k_s}|| + \cdots + ||\tilde{\gamma}_{k_D} - \tilde{\gamma}_{0,k_D}|| \prod_{s=1}^{D-1} ||\tilde{\gamma}_{0,k_s}|| \leq ||\tilde{\gamma}_{k_1} - \tilde{\gamma}_{0,k_1}|| \prod_{s=2}^{D}(||\tilde{\gamma}_{k_s} - \tilde{\gamma}_{0,k_s}|| + ||\tilde{\gamma}_{0,k_s}||) + \cdots +$ $||\tilde{\gamma}_{k_D} - \tilde{\gamma}_{0,k_D}|| \prod_{s=1}^{D-1} ||\tilde{\gamma}_{0,k_s}||$.

If $||\tilde{\gamma}_h - \tilde{\gamma}_{0,h}|| \leq u_n,\ h = 1, .., V_n$, the above inequality implies that $|\Gamma_{\boldsymbol{k}} - \Gamma_{0,\boldsymbol{k}}| \leq u_n \prod_{s=2}^{D}(u_n + ||\tilde{\gamma}_{0,k_s}||) + \cdots + u_n \prod_{s=1}^{D-1} ||\tilde{\gamma}_{0,k_s}|| \leq \upsilon_n$. Thus $\Pi(||\mathbf{\Gamma} - \mathbf{\Gamma}_0||_\infty \leq \upsilon_n) \geq \Pi(||\tilde{\gamma}_h - \tilde{\gamma}_{0,h}|| \leq u_n,\ h = 1, .., V_n)$. Therefore,

$$\Pi(\mathcal{J}|\mathcal{C}) \geq \Pi(||\tilde{\gamma}_h - \tilde{\gamma}_{0,h}|| \leq u_n,\ h = 1, .., V_n)$$
$$\geq E\left[\Pi(||\tilde{\gamma}_h - \tilde{\gamma}_{0,h}|| \leq u_n,\ h = 1, .., V_n|\boldsymbol{\zeta})\right] \geq E\left[\prod_{h=1}^{V_n} \left\{\exp\left\{-||\tilde{\gamma}_{0,h}||^2/2\right\} \Pi(||\tilde{\gamma}_h|| \leq u_n|\boldsymbol{\zeta})\right\}\right]$$
$$= \exp\left\{-\sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||^2/2\right\} E\left[\prod_{h=1}^{V_n} \Pi(||\tilde{\gamma}_{h,n}|| \leq u_n|\boldsymbol{\zeta})\right], \tag{8}$$

where the second inequality follows from Anderson Lemma. We will now make use of the fact that $\int_{-a}^{a} e^{-x^2/2}dx \geq e^{-a^2}2a$ to conclude

$$\Pi(||\tilde{\gamma}_h|| \leq u_n|\Delta) \geq \prod_{r=1}^{R_{0,n}} \Pi(|\gamma_h^{(r)}| \leq u_n/R_{0,n}|\Delta) = \prod_{r=1}^{R_{0,n}} \left((1-\Delta) + \frac{\Delta}{\sqrt{2\pi}} \int_{-u_n/R_{0,n}}^{u_n/R_{0,n}} \exp(-x^2/2)\right)$$
$$\geq \prod_{r=1}^{R_{0,n}} \left((1-\Delta) + \frac{\Delta}{\sqrt{2\pi}} \exp\left(-\frac{u_n^2}{R_{0,n}^2}\right)\left(\frac{2u_n}{R_{0,n}}\right)\right)$$
$$\geq \left[(1-\Delta) + \frac{\Delta}{\sqrt{2\pi}} \exp\left(-\frac{u_n^2}{R_{0,n}^2}\right)\left(\frac{2u_n}{R_{0,n}}\right)\right]^{R_{0,n}}.$$

$$
\begin{aligned}
\prod_{h=1}^{V_n} \Pi(||\tilde{\gamma}_h|| \le u_n) &\ge E\left[(1-\Delta) + \frac{\Delta}{\sqrt{2\pi}} \exp\left(-\frac{u_n^2}{R_{0,n}^2}\right)\left(\frac{2u_n}{R_{0,n}}\right)\right]^{R_{0,n}V_n} \\
&= E\left[\sum_{l=0}^{R_{0,n}V_n} \binom{R_{0,n}V_n}{l}(1-\Delta)^l \left(\frac{\Delta}{\sqrt{2\pi}}\right)^{R_{0,n}V_n-l} \left(\frac{2u_n}{R_{0,n}}\right)^{R_{0,n}V_n-l} \exp\left(-(R_{0,n}V_n-l)\frac{u_n^2}{R_{0,n}^2}\right)\right] \\
&\ge \left(\frac{1}{\sqrt{2\pi}}\right)^{R_{0,n}V_n} \sum_{l=0}^{R_{0,n}V_n} \binom{R_{0,n}V_n}{l} Beta(R_{0,n}V_n - l + 1, l+1) \left(\frac{2u_n}{R_{0,n}}\right)^{R_{0,n}V_n-l} \exp\left(-(R_{0,n}V_n-l)\frac{u_n^2}{R_{0,n}^2}\right) \\
&\ge \left(\frac{1}{\sqrt{2\pi}}\right)^{R_{0,n}V_n} \sum_{l=0}^{R_{0,n}V_n} \frac{(R_{0,n}V_n)!}{l!(R_{0,n}V_n-l)!} \frac{l!(R_{0,n}V_n-l)!}{(R_{0,n}V_n+1)!} \left(\frac{2u_n}{R_{0,n}}\right)^{R_{0,n}V_n-l} \exp\left(-(R_{0,n}V_n-l)\frac{u_n^2}{R_{0,n}^2}\right) \\
&\ge \left(\frac{1}{\sqrt{2\pi}}\right)^{R_{0,n}V_n} \frac{R_{0,n}V_n}{R_{0,n}V_n+1} \left(\frac{2u_n}{R_{0,n}}\right)^{R_{0,n}V_n} \exp\left(-V_n\frac{u_n^2}{R_{0,n}}\right).
\end{aligned}
$$

Aggregating all pieces together

$$
\begin{aligned}
\Pi(||\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_0||_\infty \le v_n) \ge \exp\left(-\frac{\sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||^2}{2}\right)\left(\frac{1}{\sqrt{2\pi}}\right)^{R_{0,n}V_n} \frac{R_{0,n}V_n}{R_{0,n}V_n+1}\left(\frac{2u_n}{R_{0,n}}\right)^{R_{0,n}V_n} \\
\exp\left(-V_n\frac{u_n^2}{R_{0,n}}\right).
\end{aligned}
$$

5 | APPENDIX B

Proof of Theorem 3

To begin, we define a few metrics of discrepancy between g and g_0 as below:

$$
\begin{aligned}
d_0(g, g_0) &= \int\int g_0(y|\boldsymbol{X}) \log\left(\frac{g_0(y|\boldsymbol{X})}{g(y|\boldsymbol{X})}\right) \nu_{\boldsymbol{X}}(d\boldsymbol{X})\nu_y(dy), \\
d_t(g, g_0) &= (1/t)\left\{\int\int g_0(y|\boldsymbol{X}) \left\{\frac{g_0(y|\boldsymbol{X})}{g(y|\boldsymbol{X})}\right\}^t \nu_y(dy)\nu_{\boldsymbol{X}}(d\boldsymbol{X})\right\}.
\end{aligned}
$$

For every n, define a set of probability densities given by $\mathcal{P}_n$. Let the minimum number of Hellinger balls of radius ϵ_n required to cover $\mathcal{P}_n$ be given by $\mathcal{N}_{\epsilon_n}(\mathcal{P}_n)$. To prove the theorem, it suffices to show that conditions (i)-(iii) hold:

(i) $\log \mathcal{N}_{\epsilon_n}(\mathcal{P}_n) \le n\epsilon_n^2$

(ii) $\Pi(\mathcal{P}_n^c) \le \exp(-2n\epsilon_n^2)$

(iii) For $t = 1$, $\Pi[g : d_t(g, g_0) \le \epsilon_n^2/4] \ge e^{-n\epsilon_n^2/4}$,

using Proposition 1 of Jiang, 2007. Below we show (i)-(iii) for the proposed model.

Proof of condition (ii): Define $\mathcal{P}_n$ as the set of all densities s.t. at most m_n among $\tilde{\gamma}_1, ..., \tilde{\gamma}_{V_n}$ are nonzero and each element in a nonzero $\tilde{\gamma}_h$ satisfies $|\gamma_h^{(r)}| \le C_n$. Let $g_{\boldsymbol{\zeta}}$ denote a density in $\mathcal{P}_n$ expressed with the binary variables $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_{V_n})'$. With

$|\boldsymbol{\zeta}| = \sum_{h=1}^{V_n} \zeta_h$, $\mathcal{P}_n$ contains densities $g_{\boldsymbol{\zeta}}$ s.t. $|\boldsymbol{\zeta}| \leq m_n$. Define, $\mathcal{A} = \{h \in \mathcal{N} : \zeta_h = 1\}$. Then for all large n,

$$\begin{aligned}
\Pi(\mathcal{P}_n^c) &= \Pi(|\boldsymbol{\zeta}| > m_n) + \sum_{|\boldsymbol{\zeta}| \leq m_n} \Pi(\cup_{h\in\mathcal{A}} \cup_{r=1}^{R_n} \{|\gamma_h^{(r)}| > C_n\})\Pi(\boldsymbol{\zeta}) \\
&\leq \max_{\boldsymbol{\zeta} : |\boldsymbol{\zeta}| \leq m_n} \Pi(\cup_{h\in\mathcal{A}} \cup_{r=1}^{R_n} \{|\gamma_h^{(r)}| > C_n\}) \\
&\leq R_n m_n \Pi(|\gamma_h^{(r)}| > C_n) = 2R_n m_n(1 - \Phi(C_n)) \\
&\leq \exp(\log(2R_n m_n))(1 - \Phi(C_n)) \leq exp(-2n\epsilon_n^2),
\end{aligned}$$

where the last inequality follows from assumptions (a) and (d).

Proof of condition (i): Note that, each $g_{\boldsymbol{\zeta}} \in \mathcal{P}_n$ is represented by $|\boldsymbol{\zeta}|$ nonzero $\tilde{\gamma}_h$'s with each component $\gamma_h^{(r)}$, $r = 1, ..., R_n$ of a nonzero $\tilde{\gamma}_h$ is bounded between $[-C_n, C_n]$. It takes at most $\left(1 + \frac{C_n}{\kappa}\right)^{R_n|\boldsymbol{\zeta}|}$ balls of the form $[\xi_h^{(r)} - \kappa, \xi_h^{(r)} + \kappa]$ (with their centers $\xi_h^{(r)}$'s satisfying $|\xi_h^{(r)}| \leq C_n$) to cover the parameter space of $g_{\boldsymbol{\zeta}}$. There are at most V_n^l models satisfying $|\gamma_n| = l$. Hence, the total number of balls to cover the parameter space of regression functions in $\mathcal{P}_n$ is given by $N(\kappa) = \sum_{l \leq m_n} V_n^l \left(1 + \frac{C_n}{\kappa}\right)^{R_n l} \leq (m_n + 1)\left[V_n\left(1 + \frac{C_n}{\kappa}\right)\right]^{R_n m_n}$.

Let $p_{\boldsymbol{\zeta}}$ be any density in $\mathcal{P}_n$, with $p_{\boldsymbol{\zeta}}(y|\boldsymbol{X}) = \exp(a(\mu)y + b(\mu) + c(y))$, $\mu = \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}} F_{\boldsymbol{k}}$, where $|\boldsymbol{\zeta}| \leq m_n$ and $F_{\boldsymbol{k}} = \sum_{r=1}^{R_n} \lambda_r f_{k_1}^{(r)}...f_{k_D}^{(r)}$, with $|f_h^{(r)}| \leq C_n$ for all $h \in \mathcal{A}$, $r = 1, .., R_n$. There exists a density $g_{\boldsymbol{\zeta}} \in \mathcal{P}_n$ given by $g_{\boldsymbol{\zeta}}(y|\boldsymbol{X}) = \exp(a(\alpha)y + b(\alpha) + c(y))$, with $\alpha = \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}}\Gamma_{\boldsymbol{k}}$. $\Gamma_{\boldsymbol{k}} = \sum_{r=1}^{R_n} \lambda_r \gamma_{k_1}^{(r)}...\gamma_{k_D}^{(r)}$, where $\gamma_h^{(r)}$'s are such that $f_h^{(r)} \in (\gamma_h^{(r)} - \kappa, \gamma_h^{(r)} + \kappa)$ for every r and h.

Applying Taylor expansion on $d_0(g_{\boldsymbol{\zeta}}, p_{\boldsymbol{\zeta}})$ to show that $d_0(g_{\boldsymbol{\zeta}}, p_{\boldsymbol{\zeta}}) \leq E_{\boldsymbol{X}}\left[a'(\alpha_\mu)\left(-\frac{b'(\alpha)}{a'(\alpha)}\right) + b'(\alpha_\mu)\right](\alpha - \mu)$, where α_μ is an intermediate point between α and μ. Let $\mathcal{B} = \{\boldsymbol{k} \in \mathcal{K} : \zeta_{k_1} = 1, .., \zeta_{k_D} = 1\}$. Now note that,

$$|\alpha - \mu| = |\sum_{\boldsymbol{k}\in\mathcal{B}} x_{i,\boldsymbol{k}}\Gamma_{\boldsymbol{k}} - \sum_{\boldsymbol{k}\in\mathcal{B}} x_{i,\boldsymbol{k}}F_{\boldsymbol{k}}| \leq \sum_{\boldsymbol{k}\in\mathcal{B}} |\Gamma_{\boldsymbol{k}} - F_{\boldsymbol{k}}| \leq m_n^D \max_{\boldsymbol{k}\in\mathcal{B}} |\Gamma_{\boldsymbol{k}} - F_{\boldsymbol{k}}|.$$

It follows from the above that,

$$\begin{aligned}
|\Gamma_{\boldsymbol{k}} - F_{\boldsymbol{k}}| &= |\sum_{r=1}^{R_n} \lambda_r\gamma_{k_1}^{(r)}...\gamma_{k_D}^{(r)} - \sum_{r=1}^{R_n} \lambda_r f_{k_1}^{(r)}...f_{k_D}^{(r)}| \leq |\sum_{r=1}^{R_n} \gamma_{k_1}^{(r)}...\gamma_{k_D}^{(r)} - \sum_{r=1}^{R_n} f_{k_1}^{(r)}...f_{k_D}^{(r)}| \\
&\leq \sum_{r=1}^{R_n}\left\{|\gamma_{k_1}^{(r)} - f_{k_1}^{(r)}|\prod_{l=2}^{D}|\gamma_{k_l}^{(r)}| + |f_{k_1}^{(r)}||\gamma_{k_2}^{(r)} - f_{k_2}^{(r)}|\prod_{l=3}^{D}|\gamma_{k_l}^{(r)}| + \cdots + \prod_{l=1}^{D-1}|f_{k_l}^{(r)}||\gamma_{k_D}^{(r)} - f_{k_D}^{(r)}|\right\} \\
&\leq R_n\kappa C_n^{D-1}.
\end{aligned}$$

Thus, $|\alpha - \mu| \leq m_n^D R_n\kappa C_n^{D-1}$. Similarly, $|\alpha|$, $|\mu|$ (and therefore $|\alpha_\mu|$) being bounded by $R_n C_n^D m_n^D$. Hence,

$$d_H(g_{\boldsymbol{\zeta}}, p_{\boldsymbol{\zeta}}) \leq \left\{d_0(g_{\boldsymbol{\zeta}}, p_{\boldsymbol{\zeta}})\right\}^{1/2} \leq \left\{2 \sup_{|w|\leq R_n C_n^D m_n^D} |a'(w)| \sup_{|w|\leq R_n C_n^D m_n^D} \left|\frac{b'(w)}{a'(w)}\right| \kappa R_n m_n^D C_n^{D-1}\right\}^{1/2}.$$

Choosing $\kappa = \frac{\epsilon_n^2}{2 \sup\limits_{|w| \le R_n m_n^D C_n^D} |a'(w)| \sup\limits_{|w| \le R_n m_n^D C_n^D} \left|\frac{b'(w)}{a'(w)}\right| R_n m_n^D C_n^{D-1}}$, we obtain $d_H(g_\zeta, p_\zeta) \le \epsilon_n$. Hence

$$\begin{aligned}
&\log \mathcal{N}_{\epsilon_n}(\mathcal{P}_n) \le \log N(\kappa)\\
&\le \log(m_n+1) + R_n m_n \log(V_n) + R_n m_n \log\left(1 + \frac{2 \sup\limits_{|w| \le R_n m_n^D C_n^D} |a'(w)| \sup\limits_{|w| \le R_n m_n^D C_n^D} \left|\frac{b'(w)}{a'(w)}\right| R_n m_n^D C_n^D}{\epsilon_n^2} \right)\\
&\le \log(m_n+1) + R_n m_n \log(V_n) + R_n m_n \log(2/\epsilon_n^2) + R_n m_n \log(H(R_n m_n^D C_n^D))\\
&\le n\epsilon_n^2, \quad \text{for large } n \text{, by assumptions (a)-(c).}
\end{aligned}$$

Proof of Condition (iii): Using the mean value theorem, there exists v such that $d_t(g, g_0) = E_{\boldsymbol{X}} \left\{ g'(v)(\alpha - \alpha_0) \right\}$, where $g'(\cdot)$ represents the continuous derivative function of g in the neighborhood of g_0. Let $\tau_n = \frac{\epsilon_n^2}{8 q_n}$. If for each $\boldsymbol{k} \in \mathcal{K}$, $\Gamma_{\boldsymbol{k}} \in (\Gamma_{0,\boldsymbol{k}} - \tau_n, \Gamma_{0,\boldsymbol{k}} + \tau_n)$, then

$$|\alpha - \alpha_0| = |\sum_{\boldsymbol{k} \in \mathcal{K}} x_{\boldsymbol{k}} \Gamma_{\boldsymbol{k}} - \sum_{\boldsymbol{k} \in \mathcal{K}} x_{\boldsymbol{k}} \Gamma_{0,\boldsymbol{k}}| \le \sum_{\boldsymbol{k} \in \mathcal{B}} |\Gamma_{\boldsymbol{k}} - \Gamma_{0,\boldsymbol{k}}| \le q_n \tau_n \le \epsilon_n^2/8,$$

for large n. Again, $|v| \le |\alpha - \alpha_0| + |\alpha_0| \le q_n \tau_n + \omega_n = \epsilon_n^2/8 + \omega_n$, where $\omega_n = |\alpha_0| = |\sum_{\boldsymbol{k} \in \mathcal{K}} x_{\boldsymbol{k}} \Gamma_{0,\boldsymbol{k}}| \le \sum_{\boldsymbol{k} \in \mathcal{K}} |\Gamma_{0,\boldsymbol{k}}| \le \sum_{\boldsymbol{k} \in \mathcal{K}} ||\tilde{\gamma}_{0,k_1}|| \cdots ||\tilde{\gamma}_{0,k_D}|| \le (\sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||)^D$, which is bounded by assumption (e), for sufficiently large n. Hence $||g'(v)||$ is bounded for sufficiently large n. Thus, $d_t(g, g_0) = E_x \left\{ g(v)(\alpha - \alpha_0) \right\} \le C_0 q_n \tau_n \le \epsilon_n^2/4$ for large n, for some constant C_0.

Let $C_1 = \{\boldsymbol{\Gamma} : \Gamma_{\boldsymbol{k}} \in (\Gamma_{0,\boldsymbol{k}} - \tau_n, \Gamma_{0,\boldsymbol{k}} + \tau_n), \forall\, \boldsymbol{k} \in \mathcal{K}\}$ and $C_2 = \{\lambda_1 = 1, ..., \lambda_{R_{0,n}} = 1, \lambda_{R_{0,n}+1} = 0, .., \lambda_{R_n} = 1\}$. This implies that

$$\Pi(\{g : d_t(g, g_0) \le \epsilon_n^2/4\}) \ge \Pi(C_1 \cap C_2) = \Pi(C_2)\Pi(C_1|C_2).$$

By Lemma 1, $\Pi(C_2) \ge \frac{1}{(1+R_{0,n}^\eta)^{R_n}} R_{0,n}^{\eta(R_n - R_{0,n})}$. By Lemma 2, $-\log \Pi(C_1|C_2) = -\log \Pi(||\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_0||_\infty \le \tau_n | C_2) \le \sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||^2/2 + (R_{0,n}V_n/2)\log(2\pi) + \log(1 + (1/(R_{0,n}V_n))) + R_{0,n}V_n \log(R_{0,n}) + R_{0,n}V_n \log(1/u_n) + V_n u_n^2 / R_{0,n}$. Here u_n is the minimum of the root of the equation (7) with v_n replaced by τ_n.

Since $||\tilde{\gamma}_{0,h}|| \ge 0$, $\sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||^2 \le (\sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||)^2$ is bounded for large n, by assumption (e). By assumption (i), $R_{0,n} V_n \log(R_{0,n}) = o(n\epsilon_n^2)$ (hence $R_{0,n}V_n = o(n\epsilon_n^2)$). Using the Lagrange-Maclaurin bound on the positive root of a monic polynomial of degree D, we have $u_n \le 1 + \tau_n^{1/D}$, implying $V_n u_n^2 / R_{0,n} = o(n\epsilon_n^2)$, for all large n, by assumption (a). Using Lemma 5, $1/u_n \le (\sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||)^D/\tau_n + 1$. If $G_0 = \lim_{n \to \infty} \sum_{h=1}^{V_n} ||\tilde{\gamma}_{0,h}||$, then $R_n V_n \log(1/u_n) \le R_n V_n \log(G_0^D/\tau_n) = D R_n V_n \log(G_0) + R_n V_n \log(8 q_n) + R_n V_n \log(1/\epsilon_n^2) = o(n\epsilon_n^2)$, where the last line follows from assumptions (a) and (b).

All the aforementioned calculations yield $-\log \Pi(C_1 \cap C_2) \le n\epsilon_n^2/4$, for all large n, which implies $\Pi(\{g : d_t(g, g_0) \le \epsilon_n^2/4\}) \ge \exp(-n\epsilon_n^2/4)$ for all large n. This concludes the proof.

REFERENCES

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

Belitser, E. and Nurushev, N. (2015) Needles and straw in a haystack: robust confidence for possibly sparse sequences. *arXiv preprint arXiv:1511.01803.*

Brown, D. K., Deardorff, A. V. and Stern, R. M. (1995) Estimates of a north american free trade agreement. In *Modeling North American Economic Integration*, 59–74. Springer.

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

Castillo, I., van der Vaart, A. et al. (2012) Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, **40**, 2069–2101.

Chan, S. and Kuo, C.-C. (2005) Trilateral trade relations among china, japan and south korea: Challenges and prospects of regional economic integration. *East Asia*, **22**, 33–50.

Chiang, M.-H. (2013) The potential of china-japan-south korea free trade agreement. *East Asia*, **30**, 199–216.

Craddock, R. C., Holtzheimer III, P. E., Hu, X. P. and Mayberg, H. S. (2009) Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, **62**, 1619–1628.

Guha, S. and Rodriguez, A. (2018) Bayesian regression with undirected network predictors with an application to brain connectome data. *arXiv preprint arXiv:1803.10655*.

Guhaniyogi, R. (2017) Convergence rate of bayesian supervised tensor modeling with multiway shrinkage priors. *Journal of Multivariate Analysis*, **160**, 157–168.

Guhaniyogi, R., Qamar, S. and Dunson, D. B. (2017) Bayesian tensor regression. *Journal of Machine Learning Research*, **18**, 1–31.

Hoff, P. D. (2005) Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, **100**, 286–295.

Hufbauer, G. C. (2005) *NAFTA revisited: Achievements and challenges*. Peterson Institute.

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

Martin, R., Mess, R., Walker, S. G. et al. (2017) Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, **23**, 1822–1847.

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

Relión, J. D. A., Kessler, D., Levina, E., Taylor, S. F. et al. (2019) Network classification with applications to brain connectomics. *The Annals of Applied Statistics*, **13**, 1648–1677.

Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P. and Van De Ville, D. (2011) Decoding brain states from fmri connectivity graphs. *Neuroimage*, **56**, 616–626.

Song, Q. and Liang, F. (2017) Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*.

Wei, R. and Ghosal, S. (2017) Contraction properties of shrinkage priors in logistic regression. *Preprint at http://www4. stat. ncsu. edu/˜ ghoshal/papers*.

Zhou, H., Li, L. and Zhu, H. (2013) Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, **108**, 540–552.