UNIVERSITY OF CALIFORNIA
SANTA CRUZ

INFERENCE AND UNCERTAINTY QUANTIFICATION FOR HIGH-DIMENSIONAL TENSOR REGRESSION WITH TENSOR DECOMPOSITIONS AND BAYESIAN METHODS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Daniel Spencer

June 2020

The Dissertation of Daniel Spencer
is approved:

Rajarshi Guhaniyogi, Chair

Raquel Prado

Abel Rodriguez

Juhee Lee

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Inference and Uncertainty Quantification for High-Dimensional Tensor Regression with Tensor Decompositions and Bayesian Methods

by

Daniel Spencer

The recent emergence of complex datasets in various disciplines presents a pressing need to devise regression models that have tensors either as a response or as a covariate, often under assumptions of sparsity in the corresponding tensor-valued coefficients. Models that involve tensors often require special treatment in a modeling setting due to their potentially large structures and general assumptions of sparsity in regard to associations with covariates. Importantly, scenarios with small sample sizes benefit from Bayesian methods that allow for flexible model conditions while also rigorously defining the uncertainty in any conclusions drawn from a model.

We begin with general introductions to Bayesian analysis, neuroimaging, and tensor notations in the first chapter. The goal in these short overviews is not to provide a comprehensive background, but to inform the casual reader about key concepts that will be referenced throughout this dissertation. Afterwards, we proceed through new methods in Bayesian modeling of tensor-valued variables, covering three different analysis scenarios.

The goal in the second chapter is to develop a Bayesian tensor response regression in order to identify contiguous spatial regions that are associated with a given covariate. The method is then applied to detecting neuronal activation in functional magnetic resonance imaging (fMRI) experiments in the presence of tensor-valued brain images and a scalar predictor for a single subject. We propose to regress responses from all cells (called voxels in brain activation studies)

together as a tensor response on scalar predictors, accounting for the structural information inherent in the tensor response. To estimate model parameters with proper cell specific shrinkage, we propose a novel *multiway stick breaking shrinkage prior* distribution on tensor structured regression coefficients, enabling identification of cells which are related to the predictors. The major novelty of this chapter lies in the theoretical study of the contraction properties for the proposed shrinkage prior in the tensor response regression when the number of cells grows faster than the sample size. Specifically, estimates of tensor regression coefficients are shown to be asymptotically concentrated around the true sparse tensor in L_2-sense under mild assumptions. The method is then applied to a single subject within a balloon-analog risk-taking fMRI experiment to make inferences about parts of the subject's brain that are activated by a stimulus.

In the third chapter, the Bayesian tensor response regression is expanded to compare multiple subjects with multiple tensor responses per subject. This allows for inference on a tensor-valued coefficient, as well as correlations between the different tensor response groups. These two types of inference are referred to in neuroimaging as activation and connectivity, respectively. Brain activation and connectivity analyses in task-based fMRI experiments with multiple subjects are currently at the forefront of data-driven neuroscience. In such experiments, interest often lies in understanding activation of brain voxels due to external stimuli and strong association or connectivity between the measurements on a set of pre-specified groups of brain voxels, also known as regions of interest (ROI). This chapter proposes a joint Bayesian additive mixed modeling framework that simultaneously assesses brain activation and connectivity patterns from multiple subjects. In particular, fMRI measurements from each individual obtained in the form of a multi-dimensional array/tensor across time are regressed on functions of

the stimuli. A low-rank parallel factorization (PARAFAC) decomposition on the tensor regression coefficients corresponding to the stimuli to achieve parsimony. The multiway stick-breaking shrinkage priors that were developed in the first chapter are employed to infer activation patterns and associated uncertainties in each cell within the tensor responses. Further, the model introduces region specific random effects which are jointly modeled with a Bayesian Gaussian graphical prior to account for the connectivity among pairs of ROIs. Empirical investigations under various simulation studies demonstrate the effectiveness of the method as a tool to simultaneously assess brain activation and connectivity. The method is then applied to the balloon-analog risk-taking fMRI experiment across multiple subjects in order to make inference about how the brain processes risk.

In the fourth chapter, we propose a method to parsimoniously model a scalar response with a tensor-valued covariate using the Tucker tensor decomposition. This method retains the spatial relationship within a tensor-valued covariate, while reducing the number of parameters varying within the model. Simulated data is analyzed to demonstrate model effectiveness, with comparisons made to both classical and Bayesian methods. The method is then applied to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to make inferences about the effects of Alzheimer's disease on the brain and to provide a more quantitative framework on which to make diagnoses.

Finally, we conclude with a brief review of the topics covered, research in progress, and future directions for scholastic pursuit.

For Lisa, Lua, Ira, and Otto.

Acknowledgments

This work would not exist if it were not for the support of Raj Guhaniyogi and Raquel Prado, who guided me extensively through the process of research and exploration. Their patience and understanding shaped how I think through statistical problems, and taught me how to process academic works.

My wife and best friend, Lisa, supported me in countless ways through my graduate studies. She encouraged my progress, took away my stress, and cared for our children so that I could continue in my research. On many occasions, Lisa used her positivity and humor to give me perspective on my work and the necessary balance that must be maintained with pursuits outside of research. This dissertation never would have come together without her.

My children; Lua, Ira, and Otto; granted me diversions on tough days, giving me a removed perspective that enabled me to continue my studies, even when they felt as though they were at a dead end.

My mother, Laura Steen, who passed away in November 2017, had an enormous impact on this work. Without her encouragement and guidance, I likely would never have recognized the satisfaction that can be derived through the order of mathematics. Her constant encouragement spurred me to persist through the difficulties of graduate school and early parenthood.

The students and faculty of the Applied Mathematics and Statistics departments have been essential to coping with the challenges of graduate school, and I am indebted to all of them. I am particularly thankful to Daniel Kirsner, Kurtis Shuler, and Matt Heiner. Their insights and friendship served as touchstones in my academic and professional development.

I am thankful to the Consumer and Community Bank Treasury modeling group at J.P. Morgan Chase & Co. in Columbus, Ohio, where I interned for two

summers during my graduate education. They exposed me to the practicalities of statistical modeling, as well as the importance of establishing efficient task management. My mentors there, including Deven Kapadia, Carlos Cabrera, and John Stettler, profoundly affected my programming and communication style.

Finally, I am thankful for the friendship and research partnership of Dr. Mandy Mejia, who I met late in my graduate studies. She has given me essential advice on handling neuroimaging data, which had a significant impact on the analysis in the fourth chapter.

Chapter 1

Introduction

New technological advances continue to inundate industry and research professionals with more data than ever before. As a function of this proliferation, the format and structure of data also continues to change and evolve. One such data structure is that of the tensor, also known as a multidimensional array. These data expand on the notion of a vector or matrix into an arbitrary dimension, typically codifying information about a datum in its position within the tensor.

One of the first areas of research to emerge with tensor-valued data is medical imaging. Imaging scanners use a variety of methods to obtain information about different parts of a body, such as a lung or brain. These images are sometimes two-dimensional, showing a slice of a three-dimensional structure. In other cases, the images are three-dimensional, capturing information throughout an entire organ or body part. Accurate analytical techniques for these types of data are useful for quantifying medical diagnosis and treatment.

In the following chapters, three different methods for analyzing such tensor-valued data will be presented, along with results from applications to neuroimaging studies. However, in order to improve readability, general introductions on neuroimaging data and tensor notation will be presented within this chapter.

1.1 A Crash Course in Neuroimaging

In this dissertation, a fair amount of time will be spent discussing applications to neuroimaging datasets. As such, much of the inference, even within simulation studies, may rely on some familiarity with some basic neuroscience concepts and vocabulary. As the focus will be on image analyses, most of this introduction will be centered around discussion of magnetic resonance images (MRIs) and functional magnetic resonance images (fMRIs).

MRIs are performed using large magnetic resonance scanners, which pass strong magnetic fields through a person's body. These fields have an effect on the directions that electrons in certain atoms within the body spin, and when large enough groups of atoms exhibit the same spin direction, a radiofrequency signal is generated, which can be detected by antennae placed close to the body. Hydrogen atoms are particularly good for having their spin changed, and since they are a key part of both water and fat, they can relate a fair amount of information about what is going on inside a human body (Brown et al., 2014). Since the brain is mostly made of water and fat, magnetic resonance is particularly effective at capturing images of the brain.

These images are received from the scanner after scanner-specific corrections are made by technicians. In the case of an MRI, a single three-dimensional array is returned. In all of the MRI scans used in analyses in this dissertation, T1-weighted scans are used, which present high contrasts within fats. These types of scans are well-suited to detecting different physical structures within the brain, and they are sometimes called T1-weighted structural scans. These images are in greyscale, the brightness of each volumetric pixel, or *voxel* (Lazar, 2008), a relative measure of the resonance picked up by the antennae within the scanner. An example of a T1-weighted scan from a subject within the Alzheimer's Disease Neuroimaging

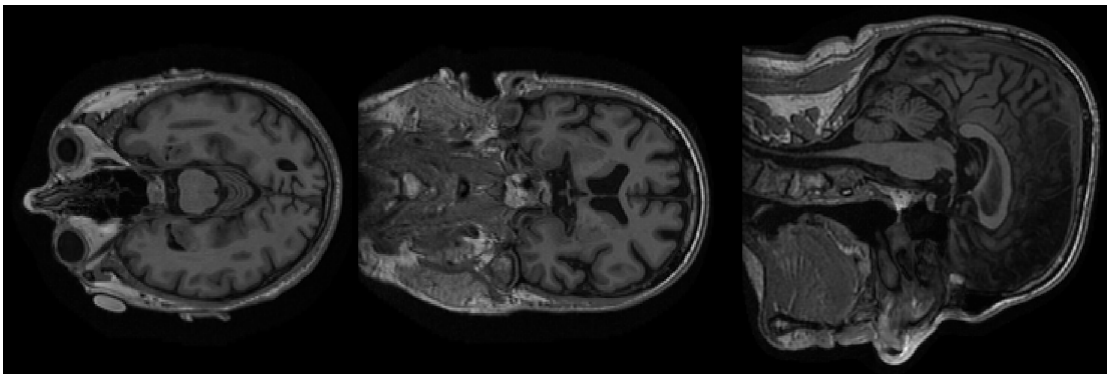

Figure 1.1: An example of a T1-weighted image from the Alzheimer's Disease Neuroimaging Initiative (`adni.loni.usc.edu`) before any preprocessing is applied.

Initiative (ADNI) image repository (`adni.loni.usc.edu`) can be seen in figure 1.1. These scans contain information about all of the structures within the head, such as the eyes, skull, mouth, and neck. Before any of these scans may be used in an analysis, several processing steps must be performed in order to be able to make accurate inference regarding brain structures or functions. In a structural scan, the first such step is to perform brain extraction, which removes all structures outside of the brain from a given image (Smith, 2002b). If multiple images from the same subject or images from multiple subjects are going to be used, registration needs to be performed in order to ensure that the same voxels from different images correspond to the same locations. In the case when a single subject is being analyzed, all T1-weighted images are typically registered to occupy the same space of a single image through a linear transformation (movement and rotation) or nonlinear transformation (movement, rotation, and scaling). When multiple subjects are within the same analysis, they are typically registered to a standard template. Standard templates, such as the MNI152 standard template (Grabner et al., 2006), are created by linearly coregistering the scans of multiple people, 152 in the case of the MNI152 standard, to create a standard space, which

can be used to compare results across multiple analyses.

In an fMRI scan, a four-dimensional array is output, three dimensions representing physical space, and a fourth dimension representing time. Functional scans measure something called the blood-oxygen-level dependent (BOLD), which measures changes in oxygen saturation using the magnetism of hemoglobin. These scans tend to be lower in resolution as they come out of the scanner, that is, before any preprocessing is performed. An example of an fMRI scan at a fixed time after brain extraction was performed is shown in figure 1.2 (Schonberg et al., 2012). Scans of the brain volume are taken around every two seconds, resulting in a time series measurement for each voxel within the scan. These scans also contain information about an entire part of a person's body, which must be processed before analysis. In addition to brain extraction, multiple steps must be taken to create analysis-ready image data (Sweeney et al., 2014). The images are motion-corrected to adjust for movements resulting from heart beats, breathing, and small conscious movements. The voxel-wise time series are typically high-pass filtered to correct for low-frequency shifts in the magnetic field, and to further correct for breathing motions. Registration takes place by first registering a T1-weighted image from the subject to a standard template, then registering the fMRI scan to the T1-weighted image, and finally by combining the two operations to register the fMRI scan to a standard space. To increase the signal-to-noise ratio within the scan, spatial smoothing is applied, such as Gaussian kernel smoothing. Finally, values in the scan may be multiplied by 1 if they are within the standard template space, and 0 if they fall outside that standard template in a process referred to as *masking*. All of the previously-listed preprocessing steps are implemented in multiple software packages, such as FSL (Smith, 2002b; Smith et al., 2004) and FreeSurfer (`surfer.nmr.mgh.harvard.edu`). Additional processing may be

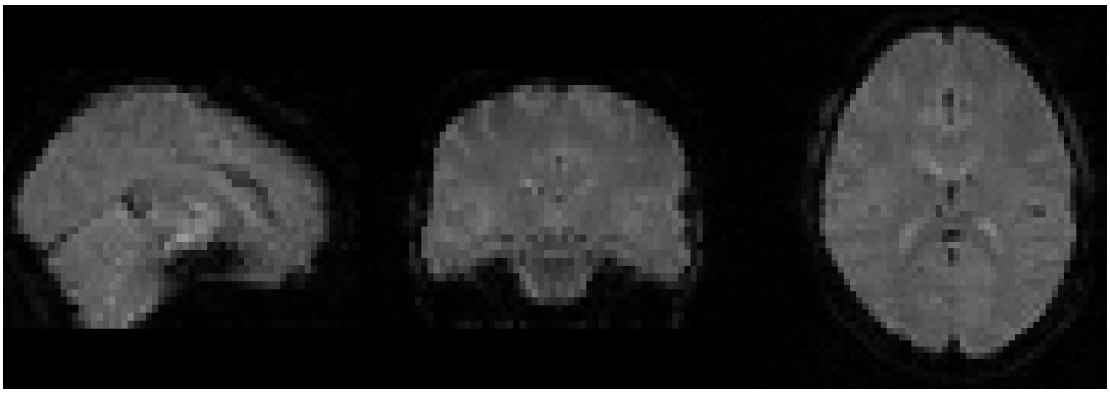

Figure 1.2: An example of the low resolution seen in fMRI data.

performed to facilitate computation. For example, in a standard space there is typically a large volume around the brain within the image that contains only zeros in the array. Occasionally, this empty space may be trimmed off in order to speed up any computation that depends on the image. Once the image data are preprocessed, the array dimensions for the scan vary widely, depending on the standard space that the data are registered to. The MNI152 template at a resolution of $2mm^3$ represents a three-dimensional volume in a tensor with dimensions $91 \times 109 \times 91$, which has 902,629 voxels. Note that in an fMRI scan registered to the MNI152 template, the number of elements in the tensor would be $902,629 \times T$, where T is the number of time steps. The number of time steps is the length of the scan, in seconds, divided by the repetition time, often denoted as TR. The TR is the number of seconds between each brain volume image, and is typically around 2 seconds. Therefore, in a 10-minute fMRI session within an experiment, T is usually around 300.

fMRI studies are centered around either task-related or resting-state scans, depending on whether subjects are instructed to receive stimuli during the scan or not. We will focus on the analysis of task-related fMRI here and in future chapters. An important consideration to take into account when analyzing fMRI data when

the subjects are completing any kind of task is the delay between the presentation of a stimulus and the physiological response observed in the scan. There are several functions that are used to apply this delay to a covariate that changes through time in an experiment. These functions are then convolved with the values of the event-related covariates. Collectively known as haemodynamic response functions (HRFs), they vary in their parameterization, affecting the shape and scale of the function. The most commonly-used HRF is referred to as the canonical, or double-gamma, HRF. The canonical HRF is characterized via six parameters, with default values given in the `neuRosim` package in R (Welvaert et al., 2011) shown in parentheses assigned as follows: delay of response relative to onset (6), delay of undershoot relative to onset (12), dispersion of response (0.9), dispersion of undershoot (0.9), scale of undershoot (0.35), and amplitude (1). A plot of the function with these default parameters can be seen in figure 1.3. While there are bodies of work surrounding the use of different HRFs and proper modeling of their parameters (Lindquist et al., 2009; Gössl et al., 2001; Marrelec et al., 2003), for the work done in the following chapters we will use the canonical HRF with the default values for the parameters. While the differences in the physiological responses may have an effect on inference, the treatment of these differences will be explored in future research.

Single-subject neuroimaging studies tend to be simpler, as they do not require registration to a standard template, and they can provide helpful insight into a subject's neurological function and health. However, many researchers conduct multiple-subject studies with the goal of learning about brain mechanics that are common to everyone within a population. The size of image data can be prohibitive in such studies, in which computing resources do not allow for the accurate, simultaneous analysis of data from dozens or hundreds of subjects. Several

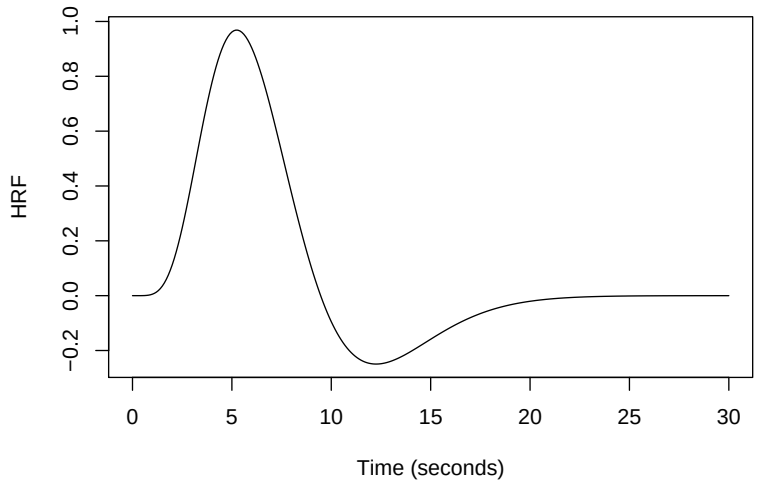

Figure 1.3: The canonical (double-gamma) haemodynamic response function

strategies have been proposed, which will be explored in later chapters.

1.2 An Introduction to Tensor Notation

Before proceeding to the next several chapters, the notation and conventions that will be used for tensor representations will be briefly explained. The term *tensor* is a generalization of an arrayed data structure. A tensor of order D is a multi-dimensional array data structure $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$. Therefore, a vector is a tensor of order 1, a matrix is a tensor of order 2, a box-shaped array is a tensor of order 3, and so forth.

The *vectorization* of a tensor $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D} : D \geq 2$ results in a tensor of order 1 of length $\prod_{j=1}^{D} p_j$, that is $\text{vec}\mathbf{B} \in \mathbb{R}^{p_1 \cdots p_D}$. The *inner product* of two tensors $\mathbf{A}$ and $\mathbf{B}$ is the crossproduct of the vectorized elements of the tensors, that is $\langle \mathbf{A}, \mathbf{B} \rangle = (\text{vec}\mathbf{A})^T (\text{vec}\mathbf{B})$.

The kth-mode *matricization* of a tensor, represented as $\mathbf{B}_{(k)}$ is a matrix representation of a tensor of order 2 or higher such that the kth index becomes the

first index and all other tensor indices are combined in order into a second index. That is, $\mathbf{B}_{(k)} \in \mathbb{R}^{p_k \times p_1 \cdots \times p_{k-1} \times p_{k+1} \times \cdots \times p_D}$.

Let $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p_1})'$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p_2})'$ be $p_1 \times 1$ and $p_2 \times 1$ vectors, respectively. The vector outer product $\boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2$ is a $p_1 \times p_2$ array with (i, j)-th entry $\beta_{1i}\, \beta_{2j}$. A D-way outer product between vectors $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})$, $1 \leq j \leq D$, is a $p_1 \times \cdots \times p_D$ dimensional array denoted by $\mathbf{B} = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \cdots \circ \boldsymbol{\beta}_D$ with entries $\mathbf{B}_{i_1,\ldots,i_D} = \prod_{j=1}^{D} \beta_{ji_j}$. Define a $vec(\mathbf{B})$ operator as one that stacks elements of this tensor into a column vector of length $\prod_{j=1}^{D} p_j$. From the definition of outer products, it follows that $vec(\boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \cdots \circ \boldsymbol{\beta}_D) = \boldsymbol{\beta}_D \otimes \cdots \otimes \boldsymbol{\beta}_1$, where $\otimes$ represents the Kronecker product. A tensor $\mathbf{B} \in \otimes_{j=1}^{D} \mathbb{R}^{p_j}$ is known as a D-way tensor. A mode-k *fiber* of a D-way tensor is obtained by fixing all dimensions of a tensor except the k-th one. For example, in a matrix (equivalently a 2-way tensor), a column is a mode-1 fiber and a row is a mode-2 fiber. A k-th mode vector product of a D-way tensor $\mathbf{B}$ and vector $\boldsymbol{a} \in \mathbb{R}^{p_k}$, denoted by $\mathbf{B} \bar{\times}_k \boldsymbol{a}$, is a tensor of the order of $p_1 \times \cdots \times p_{k-1} \times p_{k+1} \times \cdots \times p_D$, whose elements are the inner product of each mode-k fiber of $\mathbf{B}$ with $\boldsymbol{a}$.

Finally, tensors can be represented via different *tensor decompositions*. A common decomposition that is currently in use is the canonical decomposition/parallel factorization, also known as CANDECOMP/PARAFAC, or CP Tucker (1966). This decomposition represents the tensor $\mathbf{B}$ as

$$\mathbf{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_{1,r} \circ \cdots \circ \boldsymbol{\beta}_{D,r}, \tag{1.1}$$

in which $\boldsymbol{\beta}_{j,r} \in \mathbb{R}^{p_j}$ is one of R principal components for the jth dimension of $\mathbf{B}$. Here, R is known as the *rank* of the CP decomposition. The $\circ$ operator denotes the *outer product*. At each value of $r \in \{1, \ldots, R\}$, the series of outer products

$\boldsymbol{\beta}_{1,r} \circ \cdots \circ \boldsymbol{\beta}_{D,r}$ results in a D-dimensional tensor summand $\mathbf{B}_r \in \mathbb{R}^{p_1,\ldots,p_D}$. As the value of R increases, the resolution allowed by the CP decomposition improves. All tensors can conform to this CP decomposition for some appropriate rank R. In practice, low-rank decompositions have been found to be adequate to estimate sparse, smooth tensor coefficients. Using the CP decomposition reduces the parameter space for estimating $\mathbf{B}$ from $\prod_{j=1}^{D} p_j$ to $R\sum_{j=1}^{D} p_j$, which results in better accuracy detecting signal that is sparse and in contiguous box-shaped groupings. However, it is possible that not all dimensions of a tensor require all R principal components in order to be faithfully represented by such a decomposition. In order to address such a situation, the CP decomposition is extended to the *Tucker decomposition*, which can be written as

$$\mathbf{B} = \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D} g_{r_1,\ldots,r_D} \boldsymbol{\beta}_{1,r_1} \circ \cdots \circ \boldsymbol{\beta}_{D,r_D}, \tag{1.2}$$

where

$$(g_{1,\ldots,1}, g_{2,\ldots,1}, \ldots, g_{R_1,\ldots,R_D}) = \mathbf{G} \in \mathbb{R}^{R_1,\ldots,R_D}$$

is the *core tensor* composed of elements, which assigns weights of importance to each of the $\prod_{j=1}^{D} R_j$ tensor summands that compose the tensor $\mathbf{B}$ (Tucker, 1966). This representation is more flexible and parsimonious than the CP decomposition, as it allows for different ranks for different tensor dimension margins. The parameter space for estimating $\mathbf{B}$ can be reduced from $R\sum_{j=1}^{D} p_j$ when using the CP decomposition to $\sum_{j=1}^{D} R_j p_j$ with the Tucker decomposition.

With these basic conventions, a linear model can be built using the Tucker decomposition that is parsimonious with regard to the parameter space.

With these basic concepts in mind, we can now apply them in the analysis of

large datasets in neuroscience, starting with a tensor response regression.

Chapter 2

Bayesian Tensor Response Regression With an Application to Brain Activation Studies

We begin with a tensor response regression model for one subject in which the response and the covariate are time varying. This particular model relies on the simpler CP/PARAFAC tensor decomposition, combined with a novel prior structure. The method will be applied to simulated data to show its effectiveness under modeling assumptions, and then applied to a single subject fMRI experiment.

2.1 Introduction

Neuroscience and related imaging applications routinely encounter regression scenarios involving a multidimensional array or tensor structured response and scalar predictors. An important motivating example occurs in single-subject Functional MRI (fMRI) studies to detect localized regions where neuronal activation takes place in presence of external stimuli (e.g., during a task).

The tensor response at each time point is presumed to be associated with the task related predictors and it is of scientific interest to delineate the nature and region of activation using a regression framework involving the tensor response and task related predictors. Similarly, in electroencephalography (EEG) studies voltage values are measured from numerous electrodes placed on scalp over time. The resulting data are in a two-dimensional matrix where the readings are both spatially and temporally correlated. These matrix responses are often regressed on a set of scalar predictors (e.g. if a subject is an alcoholic or not) to identify their variation with the predictors. All these applications involve a response tensor $\mathbf{Y}_t \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ and a vector of predictors $\mathbf{x}_t \in \mathbb{R}^m$ at time t respectively. The objective in these experiments is to understand which cells in $\boldsymbol{Y}_t$ are influenced by the changes in $\mathbf{x}_t$, and by how much. Although the tensor response regression framework is motivated by aforementioned neuroimaging studies, the proposed methodology equally applies to a variety of scientific applications, including chemometrics (Bro, 2006), psychometrics (Kiers and Mechelen, 2001) and relational data (Gerard and Hoff, 2015), among others, where tensor valued responses are collected routinely.

Rather than analyzing cells in a tensor response together, the popular General Linear Model (GLM), sometimes referred to as a mass univariate analysis (MUA), fits a regression model at each cell in the tensor response independently of the others and calculates the test statistic corresponding to each cell to identify if the response is significantly associated with a predictor in that cell, accounting for multiple testing corrections (Penny et al., 2011; Friston et al., 1995; Genovese et al., 2002; Lindquist and Mejia, 2015). While the acronym GLM may cause statisticians to think of the generalized linear model, we will be referring to the General Linear Model to remain consistent with the neuroimaging literature. The

GLM is conceptually simple and computationally efficient, though it fails to accommodate spatial associations across cells in the tensor response. Additionally, neuroimaging data are usually pre-processed using a kernel convolution based spatial smoothing approach. Performing a GLM on pre-smoothed data may result in inaccurate estimation and testing of the covariate effects (Chumbley and Friston, 2009; Li et al., 2011). More principled approaches vectorize the tensor response to construct a multivariate vector response regression. Some notable structures employed to estimate parameters in the multivariate vector response regression include sparse regressions with various penalties incorporating correlated response variables (Similä and Tikka, 2007; Peng et al., 2010), reduced-rank regressions (Yuan et al., 2007; Chen et al., 2013) and sparse reduced-rank regressions (Chen and Huang, 2012). While these methods view tensor response as a high dimensional vector without any spatial association among its cells, our goal is to incorporate spatial information in the multidimensional tensor into the proposed model.

To this end, sophisticated approaches include adaptive multiscale smoothing methods and spatially varying coefficient (SVC) models. The former estimates parameters by building iteratively increasing neighbors around each cell and combining observations within the neighbors with weights (Li et al., 2011). The SVC models add spatial components in the cell by cell regression that account for the spatial correlations between cell (Zhang et al., 2015, 2014b; Descombes et al., 1998; Zhu et al., 2014). There is a parallel literature to model spatial dependence among regression coefficients induced by Markov random fields (MRF) (Smith and Fahrmeir, 2007). These approaches introduce distinct parameters for different cell specific regressions and propose to model them jointly. For a tensor response of dimensions $p_1 \times \cdots \times p_D$, where $p_1,\ldots,\ p_D$ are moderately large, such strategies

lead to the joint modeling of *at least* $\prod_{i=1}^{D} p_i$ parameters, which may turn out to be computationally challenging.

Recently, Li and Zhang (2017) propose a novel approach of regressing the tensor variate response on scalar predictors, where recently developed *envelope* technique by Cook et al. (2010) is employed to yield point estimates of the parameters. Subsequently, Sun and Li (2017) provide convergence rates of the frequentist penalized regression approaches with a tensor response and vector predictors. This approach proposes low rank decomposition of the tensor coefficient and introduces multiple constraints on the parameter space. While such constraints can be easily accommodated by frequentist optimization algorithms, they offer a steep challenge for Bayesian implementation. Additionally, frequentist optimization frameworks are dependent on tuning parameters (e.g., the envelope dimensions in Li and Zhang (2017)), with choices for these parameters being sensitive to the tensor dimensions and the signal-to-noise ratio (degree of sparsity).

In the same vein as Li and Zhang (2017), we propose a regression scenario with tensor response $\boldsymbol{Y}_t$ and predictors $\mathbf{x}_t$, referred to as the *tensor response regression* (TRR). The coefficient corresponding to each predictor in the vector $\mathbf{x}_t$ is a tensor, and is assumed to possess a "low rank" canonical decomposition/parallel factorization decomposition (CANDECOMP/PARAFAC, or CP), which is defined in Section 1.2. The model is also designed to be generalizable to any value of D for possible application in other research areas. For the Bayesian implementation, we employ a novel *multiway stick-breaking shrinkage prior* distribution to shrink the cells of the tensor coefficient corresponding to unimportant voxels close to zero while maintaining accurate estimation and uncertainty of cell coefficients related to important voxels. Our framework is, to the best of our knowledge, the first Bayesian framework for regressing a tensor response on scalar predictors. Addi-

tionally, TRR retains the tensor structure of the response to implicitly preserve correlations between cells and yet substantially reduces the number of parameters using the CP decomposition to accrue computational benefits. The TRR framework with the multiway stick-breaking prior gives rise to model-based shrinkage towards a "low rank" solution for the tensor coefficient, with a carefully constructed shrinkage prior that naturally induces sparsity within and across ranks for the tensor coefficient and results in identification of important cells in the tensor related to a predictor. In addition, there is a strong need for uncertainty quantification for parametric estimates, especially when the tensor dimension far exceeds the sample size, or the signal to noise ratio is low, motivating the Bayesian TRR (BTRR) approach.

There is a recent literature on regressing a scalar response on a tensor covariate (Guhaniyogi et al., 2017; Zhou et al., 2013; Zhou and Li, 2014) that focuses on identifying voxels in the tensor which are related to the response. In contrast, we flip the role and regress a tensor response on scalar predictors. Our approach differs from the existing frequentist and Bayesian tensor modeling approaches (Gerard and Hoff, 2015; Dunson and Xing, 2009) as we offer a supervised tensor regression framework that accommodates scalar predictors.

One important contribution of this chapter remains proving posterior consistency for the proposed BTRR model with the multiway stick-breaking shrinkage prior. Theory of posterior contraction for high dimensional regression models has gained traction lately, though the literature is less developed in shrinkage priors compared to point-mass priors. For example, Castillo et al. (2012) and Belitser and Nurushev (2015) have established posterior concentration and variable selection properties for certain point-mass priors in the many normal-means model. The latter article also establishes coverage of Bayesian credible sets. Results on

posterior concentration and variable selection in high dimensional linear models are also established by Castillo et al. (2015a) and Martin et al. (2017) for certain point-mass priors. In contrast, Armagan et al. (2013b) show posterior consistency in the linear regression model with shrinkage priors for low-dimensional settings where the number of covariates *does not* exceed the number of observations. Using direct calculations, Van Der Pas et al. (2014) show that the posterior based on the horseshoe prior concentrates at the optimal rate for the many normal-mean problem. Song and Liang (2017) and Wei and Ghosal (2017) consider a general class of continuous shrinkage priors and obtain posterior contraction rates in ordinary high dimensional linear regression models and logistic regression models respectively, depending on the concentration and tail properties of the density of the continuous shrinkage prior. In contrast, the study of posterior contraction properties for tensor regression models in the Bayesian paradigm has been given far too little attention. A recent article by Guhaniyogi (2017) is of interest in this regard. Developing theory for tensor response regression models is faced with two major challenges. While high dimensional regression models directly impose a well-investigated shrinkage prior on the predictor coefficients, BTRR imposes shrinkage priors on margins of the CP decomposition of tensor coefficients. As a result, the prior distribution on voxel level elements of the tensor coefficient is difficult to deal with. Additionally, in typical applications, the dimensions of tensor coefficients are much larger than the sample size. Both of these present obstacles which we overcome in this work. We also emphasize that the posterior contraction of tensor regression in Guhaniyogi (2017) is shown for the Kullback-Leibler neighborhood. In contrast, Bayesian tensor response regression develops a much stronger result with L_2-neighborhood around the true tensor coefficient.

The remainder of the chapter flows as follows. Section 2.2 introduces the model

and describes prior distributions on the parameters. Section 2.3 describes results on posterior consistency of the proposed model. Section 2.4 and 2.5 show performance of the proposed model through simulation studies and brain activation data analysis. Section 2.6 concludes the paper.

2.2 Framework and Model

2.2.1 Model framework

Let $\boldsymbol{Y}_t = ((Y_{t,\boldsymbol{v}}))_{v_1,...,v_D=1}^{p_1,..,p_D} \in \otimes_{j=1}^{D} \mathbb{R}^{p_j}$ denote a tensor valued response at time t, where $\boldsymbol{v} = (v_1, ..., v_D)'$ represents the position of voxel $\boldsymbol{v}$ in the D dimensional array of voxels. Let $\mathbf{x}_t = (x_{1,t}, ..., x_{m,t})' \in \mathcal{X} \subset \mathbb{R}^m$ be the m-dimensional measured vector predictor. Assuming that both response $\boldsymbol{Y}_t$ and predictors $\mathbf{x}_t$ are centered around their respective means, the proposed tensor response regression model of $\boldsymbol{Y}_t$ on $\mathbf{x}_t$ is given by

$$\boldsymbol{Y}_t = \mathbf{B}_1 x_{1,t} + \cdots + \mathbf{B}_m x_{m,t} + \boldsymbol{E}_t, \tag{2.1}$$

for $t = 1, ..., T$. $\mathbf{B}_k \in \otimes_{j=1}^{D} \mathbb{R}^{p_j}$, $k = 1, .., m$ is the tensor coefficient corresponding to the predictor $x_{k,t}$. To account for the temporal correlation of the response tensor, the error tensor $\boldsymbol{E}_t \in \otimes_{j=1}^{D} \mathbb{R}^{p_j}$ is assumed to follow a componentwise AR(1) structure, $\boldsymbol{E}_t = \kappa \boldsymbol{E}_{t-1} + \boldsymbol{\nu}_t$, where $\kappa \in (-1, 1)$ is the autocorrelation coefficient and $\boldsymbol{\nu}_t \in \otimes_{j=1}^{D} \mathbb{R}^{p_j}$ with each cell in $\boldsymbol{\nu}_t$ following $N(0, \sigma^2/(1-\kappa^2))$. This ensures both computational simplicity and stationarity in the AR(1) structure.

Naive voxel-by-voxel regression of $Y_{t,v}$ on $\mathbf{x}_t$ requires introducing m regression parameters per voxel, hence a total of $m \prod_{j=1}^{D} p_j$ parameters, resulting in an ultra-high dimensional modeling pursuit, and fails to incorporate tensor struc-

tural information into the estimation procedure. This necessitates imposing a sufficiently expressive structure on $\mathbf{B}_k$ which simultaneously achieves a large dimensionality reduction. We propose flexible rank-R CP decomposition of each $\mathbf{B}_k$, i.e. $\mathbf{B}_k = \sum_{r=1}^{R} \boldsymbol{\beta}_{1,r,k} \circ \cdots \circ \boldsymbol{\beta}_{D,r,k}$, where $\boldsymbol{\beta}_{j,r,k} = (\boldsymbol{\beta}_{j,r,k,1}, ..., \boldsymbol{\beta}_{j,r,k,p_j})'$ is a p_j dimensional vector, $1 \leq r \leq R$, $1 \leq j \leq D$ and $k = 1, .., m$.

A few remarks on (2.1) are in order. First, since we deal with modeling the linear predictor part of the model, our framework can be extended to a GLM set up. Second, the formulation also assumes easy extensions to settings with a more complicated spatio-temporal correlation structure in $\boldsymbol{E}_t$. Additionally, CP decomposition reveals that the cell-level parameters are nonlinear functions of the tensor margins $\boldsymbol{\beta}_{j,r,k}$. Careful choice of prior distributions on the tensor margins implicitly imposes correlations among voxels and facilitates identifying significantly nonzero cells in $\mathbf{B}_k$.

Imposing this additional rank-R CP structure on $\mathbf{B}_k$ remarkably reduces the total number of parameters in the model from $m \prod_{j=1}^{D} p_j$ to $Rm \sum_{j=1}^{D} p_j$. A critical question remains whether such a dimension reduced structure can identify geometric sub-regions in the tensor response which are related to the predictors. Additionally, we also intend to accurately estimate coefficients corresponding to these sub-regions of the tensor coefficient. The next section proposes a careful elicitation of the prior distribution on the tensor parameters to achieve this goal.

2.2.2 Multiway stick-breaking shrinkage prior on tensor coefficients

Although the spike-and-slab prior for selective predictor inclusion (George and McCulloch, 1993; Clyde et al., 1996) possesses attractive theoretical properties, intractability of exploring an exponentially large space of predictor inclusion along

with the belief that many regression coefficients may be small rather than exactly zero has led to considerable growth in the appeal for continuous shrinkage priors. An impressive variety of Bayesian shrinkage priors for ordinary high dimensional regression with a scalar/vector response on high dimensional vector predictors has been proposed in recent times, see for example Hans (2009); Carvalho et al. (2010); Armagan et al. (2013a) and references therein. Shrinkage priors are based on the principle of artfully shrinking predictor coefficients of unimportant predictors to zero, while maintaining proper estimation and uncertainty of the important predictor coefficients. Polson and Scott (2010) further show that most of the existing shrinkage priors can be expressed as the scale mixture of normal distributions with a global parameter common to all predictors and predictor-specific local parameters. The global parameter imposes shrinkage globally while local parameters carefully balance shrinkage for large and small coefficients.

Literature on the vector shrinkage priors provides an excellent starting point for studying multiway shrinkage priors on tensor coefficient $\mathbf{B}_k$, though the latter presents a lot more challenges. Assuming that $\mathbf{B}_k$ admits a rank-R CP decomposition, proposing a prior on $\mathbf{B}_k$ is equivalent to specifying priors over tensor margins $\boldsymbol{\beta}_{j,r,k}$. Given that every cell coefficient in $\mathbf{B}_k$ is a nonlinear function of the tensor margins, care should be taken while imposing prior shrinkage on them. To this end, Guhaniyogi et al. (2017) have characterized multiple restrictions on putting prior distributions on $\mathbf{B}_k$'s and have proposed the multiway dirichlet generalized double pareto (M-DGDP) shrinkage prior satisfying all the restrictions. However, in the context of BTRR, a straightforward application of M-DGDP prior on $\mathbf{B}_k$ leads to inaccurate estimation due to less desirable tail behavior of the distribution of $B_{\boldsymbol{v},k}$ parameters.

Norm-based penalizations, such as the l_0-norm regularized least squares re-

gression model (Polson and Sun, 2019), may be considered as alternatives to continuous shrinkage priors like the generalized double-Pareto prior. Norm-penalized variable selections rely on assumptions of relatively large signal-to-noise ratios and scale to a few hundred different parameter values through the use of greedy stepwise variable selection algorithms. These assumptions are violated in many neuroimaging applications, with signal-to-noise and contrast-to-noise ratios often less than 1 (Welvaert and Rosseel, 2013), and predictors numbering in the thousands and tens of thousands.

This chapter proposes a multiway stick-breaking shrinkage prior on $\mathbf{B}_k$ to ensure desirable tail behavior for the tensor coefficient. More specifically, set $\tau_{r,k} = \phi_{r,k}\tau_k$, as the scaling specific to rank $r = 1, ..., R$. To achieve effective shrinkage across ranks we adopt a stick-breaking construction for the rank-specific scale parameters $\phi_{r,k}$s, $\phi_{r,k} = \xi_{r,k}\prod_{l=1}^{r-1}(1-\xi_{l,k})$, $r = 1, ..., R-1$, and $\phi_{R,k} = \prod_{l=1}^{R-1}(1-\xi_{l,k})$, where $\xi_{r,k} \overset{iid}{\sim} Beta(1, \alpha_k)$. The global scale parameter is modeled as $\tau_k \sim \text{Gamma}(a_\tau, b_\tau)$. Additionally, the local scale parameters $\boldsymbol{W}_{j,r,k} = \text{diag}(w_{j,r,k,1}, ..., w_{j,r,k,p_j})$ are employed to achieve margin level shrinkage in the following way

$$\begin{aligned}
\boldsymbol{\beta}_{j,r,k} &\sim \text{N}(\mathbf{0}, \tau_{r,k}\boldsymbol{W}_{j,r,k}), \\
w_{j,r,k,\ell} &\sim \text{Exp}(\lambda_{j,r,k}^2/2), \\
\lambda_{j,r,k} &\sim \text{Gamma}(a_\lambda, b_\lambda), \\
\ell &= 1, ..., p_j.
\end{aligned}$$

The construction tacitly exploits the finite stick-breaking construction for the local parameters $\phi_{r,k}$'s. As $\alpha_k \to 0$, most of the $\phi_{r,k}$s become more sparse. Therefore, careful learning of α_k leads to a sparse and parsimonious representation of

the tensor. α_k is assigned a discrete uniform prior on a grid and is learned using a griddy-Gibbs algorithm. Additionally, flexibility in estimating tensor margins $\{\boldsymbol{\beta}_{j,r,k} : 1 \leq j \leq D,\ 1 \leq r \leq R\}$ is accommodated by modeling heterogeneity within margins via element-specific scaling $\boldsymbol{W}_{j,r,k}$. A common rate parameter $\lambda_{j,r,k}$ encourages sharing of information between the margin elements. In fact, it is easy to see that $\beta_{j,r,k,\ell}|\phi_{r,k}, \tau_k$ follows the well known generalized double Pareto (GDP) (Armagan et al., 2013a) shrinkage prior distribution. Exploiting more efficient computational techniques, TRR with the multiway stick-breaking shrinkage prior accurately estimates the posterior distribution of $\mathbf{B}_k$ for a relatively large number of cells compared to the ordinary spike-and-slab prior on cell coefficients.

Under a Bayesian framework, parameter estimation can be achieved via Markov chain Monte Carlo (MCMC) algorithms, in which posterior distributions for the unknown quantities are approximated with empirical distributions of samples from a Markov chain. The full conditional distributions for developing Metropolis within Gibbs sampling algorithms are provided in appendix A.2.

2.3 Posterior consistency in tensor response regression

2.3.1 Notations

In what follows, we add a subscript (T) to the dimensions of tensor margins $p_{1,(T)}, ..., p_{D,(T)}$ and the number of predictors $m_{(T)}$ to indicate that the size of both the response tensor $\boldsymbol{Y}_t$ and covariates $\mathbf{x}_t$ can increase with the sample size T. This asymptotic paradigm is also meant to capture the fact that the number of cells $\prod_{j=1}^{D} p_{j,(T)}$ is typically larger than the sample size T for the tensor coefficients $\mathbf{B}_{1,(T)}, .., \mathbf{B}_{m_{(T)},(T)}$. Define $\mathbf{B}$ as an $\mathbb{R}^m \otimes_{j=1}^{D} \mathbb{R}^{p_j}$ tensor with the

$(v_1, .., v_D, k)$th cell being given by the $(v_1, ..., v_D)$th cell of $\mathbf{B}_{k,(T)}$. Naturally, the tensor coefficient $\mathbf{B}$ and tensor margins $\boldsymbol{\beta}_{j,r,k}$s are also functions of the sample size T and we denote them by $\mathbf{B}_{(T)}$ and $\boldsymbol{\beta}_{j,r,k,(T)}$s respectively. We use superscript (0) to indicate true parameters, e.g. the true tensor regression parameter and the true error variance are denoted by $\mathbf{B}^{(0)}_{(T)}$ and $\sigma^{(0)2}$ respectively. For simplicity, we assume that $\sigma^2 = \sigma^{(0)2}$ is known and fixed at 1. We also assume that κ is fixed and known, so that $var(\boldsymbol{E_v}) = \mathbf{S}$ is fixed, where $\boldsymbol{E_v} = (E_{1,\boldsymbol{v}}, ..., E_{T,\boldsymbol{v}})'$. While κ and σ^2 are unknown in practice and are assigned prior distributions, our setup assumes them to be fixed and known. This is a common assumption in the asymptotic study (Van der Vaart and Van Zanten, 2011). Furthermore, it is known that the theoretical results obtained by assuming these parameters as known constants are equivalent to those obtained by assigning priors with bounded supports on these parameters (Van der Vaart and Van Zanten, 2009). For vectors, we let $||\cdot||_2$ denote the L_2-norm, $||\cdot||_1$ denote the L_1-norm and $||\cdot||_\infty$ denote the L_∞ norm. With a slight abuse of notations, for a D-dimensional tensor object $\boldsymbol{A}$, the L_1, L_2 and L_∞ norms are defined as $||\boldsymbol{A}||_1 = \sum_{v_1,...,v_D} |A_{v_1,...,v_D}|$, $||\boldsymbol{A}||_2 = \sqrt{\sum_{v_1,...,v_D} A^2_{v_1,...,v_D}}$ and $||\boldsymbol{A}||_\infty = \max_{v_1,..,v_D} |A_{v_1,...,v_D}|$. $||\cdot||_0$ denotes the L_0-norm, i.e. the number of non-zero entries, for both vectors and tensors. Further, assume $\mathcal{F}_1 = \Big\{\boldsymbol{h}_1 = (v_1, ..., v_D) : 1 \le v_1 \le p_{1,(T)}, ..., 1 \le v_D \le p_{D,(T)}\Big\}$, $\mathcal{F}_2 = \Big\{h_2 = v_{D+1} : 1 \le v_{D+1} \le m_{(T)}\Big\}$. Denote $\zeta^{(0)} = \{(\boldsymbol{h}_1, h_2) : B^{(0)}_{\boldsymbol{h}_1,h_2,(T)} \neq 0, \boldsymbol{h}_1 \in \mathcal{F}_1, h_2 \in \mathcal{F}_2\}$ as a set of indices corresponding to the nonzero cells of the true tensor coefficient, and also denote $\zeta^{(0)}_1 = \{\boldsymbol{h}_1 \in \mathcal{F}_1 : B^{(0)}_{\boldsymbol{h}_1,h_2,(T)} \neq 0, \text{for some}\, h_2 \in \mathcal{F}_2\}$. Similarly, for any set $\zeta \subseteq \mathcal{F}_1 \times \mathcal{F}_2$, define $\zeta_1 = \{\boldsymbol{h}_1 \in \mathcal{F}_1 : (\boldsymbol{h}_1, h_2) \in \zeta\}$ and $\zeta_{2,\boldsymbol{h}_1} = \{h_2 \in \mathcal{F}_2 : (\boldsymbol{h}_1, h_2) \in \zeta\}$. $|\zeta|$ denotes the cardinality of the set ζ. We let $s_{(T)}$ (dependent on T) denote the number of nonzero entries in the true tensor coefficient, i.e., $s_{(T)} = ||\mathbf{B}^{(0)}_{(T)}||_0$. Let $e_{max}(\cdot)$ and $e_{min}(\cdot)$ denote the largest and

smallest eigenvalues of a square matrix, respectively.

Since the shrinkage prior on $\mathbf{B}_{(T)}$ assigns zero probability at the point zero, the exact number of nonzero elements of $\mathbf{B}_{(T)}$ is always $m_{(T)} \prod_{j=1}^{D} p_{j,(T)}$. A meaningful comparison with the value $s_{(T)}$ is made by considering $\tilde{s}_{(T)}$, the number of elements of $\mathbf{B}_{(T)}$ exceeding in absolute value a threshold a_T, which will be specified later. In other words, only elements with absolute value larger than a_T will be treated as significant and counted towards non-zero entries.

Define $\mathcal{B}_T = \Big\{\text{At least } \tilde{s}_{(T)} \text{ absolute values of } \mathbf{B}_{(T)} \text{ are greater than } a_T\Big\}$, $\mathcal{C}_T = \Big\{\mathbf{B}_{(T)} : ||\mathbf{B}_{(T)} - \mathbf{B}_{(T)}^{(0)}||_2 > \epsilon\Big\}$ and $\mathcal{A}_T = \mathcal{B}_T \cup \mathcal{C}_T$. Further suppose $\pi_T(\cdot)$ and $\Pi_T(\cdot)$ are the prior and posterior densities of $\mathbf{B}_{(T)}$ with T observations, so that

$$\Pi_T(\mathcal{A}_T) = \frac{\int_{\mathcal{A}_T} f(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T|\mathbf{B}_T)\pi_T(\mathbf{B}_T)}{\int f(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T|\mathbf{B}_T)\pi_T(\mathbf{B}_T)},$$

where $f(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T|\mathbf{B}_{(T)})$ is the joint density of $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T$ under model (2.1). This chapter intends to show

$$\Pi_T(\mathcal{A}_T) \to 0, \text{ a.s., when } T \to \infty. \tag{2.2}$$

2.3.2 Main results

The following theorem shows that (2.2) holds under mild sufficient conditions on $s_{(T)}$, $\tilde{s}_{(T)}$ and $p_{j,(T)}$s. The proof of the theorem is given in the appendix.

Theorem 2.3.1. *Denote* $p_{(T)} = m_{(T)} \prod_{j=1}^{D} p_{j,(T)}$. *Let*

(a) $\mathbf{B}_{k,(T)}^{(0)}$ *assumes a rank-*R_0 *CP decomposition,* $\mathbf{B}_{k,(T)}^{(0)} = \sum_{r=1}^{R_0} \boldsymbol{\beta}_{1,r,k,(T)}^{(0)} \circ \cdots \circ \boldsymbol{\beta}_{D,r,k,(T)}^{(0)}$, *for* $k = 1, .., m_{(T)}$, *with* $R > R_0$ *and* $||\boldsymbol{\beta}_{j,r,k,(T)}^{(0)}|| < \infty$;

(b) $||\mathbf{B}_{k,(T)}^{(0)}||_0 = s_{(T)}$, *with* $s_{(T)} \log(p_{(T)}) = o(T)$;

(c) $\tilde{s}_{(T)} \log(p_{(T)}) = O(T)$*;*

(d) $m_{(T)} \sum_{j=1}^{D} p_{j,(T)} \log(p_{j,(T)}) = o(T)$*;*

(e) There exists $\lambda_0, \lambda_1 > 0$ *s.t.* $e_{min}(\mathbf{X}'_{\nabla}\mathbf{S}^{-1}\mathbf{X}_{\nabla}) \geq T\lambda_0^2$ *and* $e_{max}(\mathbf{X}'_{\nabla}\mathbf{S}^{-1}\mathbf{X}_{\nabla}) \leq T\lambda_1^2$*, for any set* $\nabla \subseteq \{1, ..., m_{(T)}\}$*, where* $\mathbf{X}_{\nabla}$ *is a submatrix of* $\mathbf{X} = [\mathbf{x}'_1 : \cdots : \mathbf{x}'_T]'$ *with columns corresponding to the indices* ∇*.*

Under conditions (a)-(e), (2.2) holds with $a_T = \frac{\epsilon}{2p_{(T)}}$*.*

Remark: Condition (a) in Theorem 2.3.1 assumes a low-rank decomposition for the true tensor coefficient. This is a mild condition as most applications allow low-rank structure for the true tensor coefficients. Regarding condition (b), note that $s_{(T)}$ is the sparsity of the true tensor and $p_{(T)}$ is the total number of cells in the tensor. When the tensor is just a scalar ($D = 0$), i.e., the tensor regression reduces to an ordinary high dimensional regression with $m_{(T)}$ predictors, the condition reduces to $s_{(T)} \log(m_{(T)}) = o(T)$, which is a typical assumption in ordinary high dimensional regression, see Song and Liang (2017). Condition (c) also assumes the same condition for the "near sparsity" in the estimated $\mathbf{B}_{(T)}$ in the sense of $\mathcal{B}_T$. Condition (d) in Theorem 2.3.1 requires that $m_{(T)} \sum_{j=1}^{D} p_{j,(T)}$ grows sub-linearly with sample size T. However, the number of cells $m_{(T)} \prod_{j=1}^{D} p_{j,(T)}$ in the tensor $\mathbf{B}_{(T)}$ can grow at a rate much faster than the sample size T; hence, the modeling framework allows large tensor responses even for moderate sample sizes. Condition (e) is equivalent to a lower bounded compatibility number condition assumed in the theoretical study of ordinary high dimensional regression, (see Song and Liang (2017); Castillo et al. (2015b)). Finally, condition (e) also ensures $e_{max}(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})$ grows sub-linearly with T.

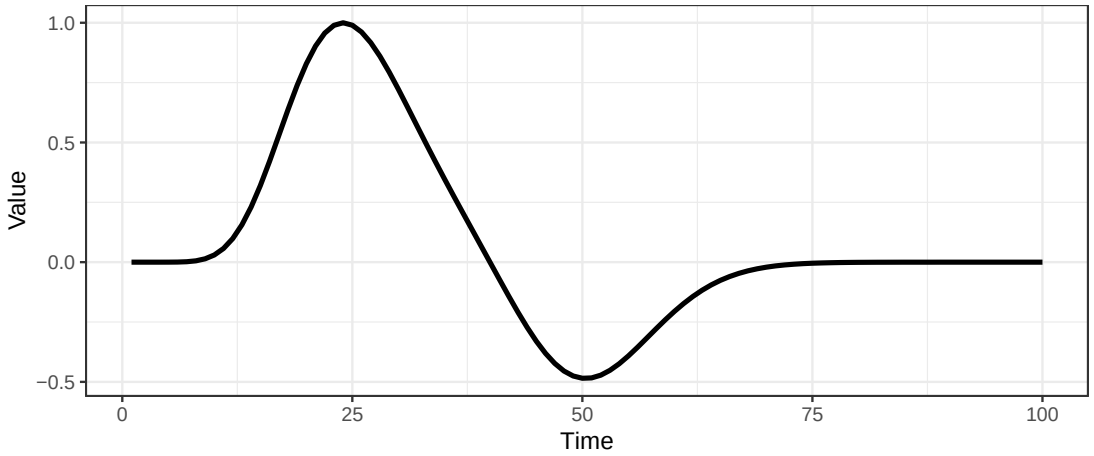

Figure 2.1: Values taken by the simulated covariate through time when the number of total time steps was set to $T = 100$.

2.4 Simulated Data Results

This section showcases parametric inference from Bayesian tensor response regression (BTRR) with various simulation studies. Since the major motivation of model development is drawn from the fMRI based brain activation study, the simulation study is performed on simulated datasets reminiscent of real world fMRI data. Scalar predictors are simulated with the block experimental design. A single stimulus block was convolved with the canonical double-gamma haemodynamic response function. An example of the values taken by the covariate can be seen in figure 2.1. A more thorough discussion on how the covariate values were generated can be found below.

The block design consisted of a single discrete epoch of activity and rest, with the “activity" representing a period of stimulus presentation, and the “rest" referring to a state of rest or baseline. The stimulus was assumed to take place at time $t = 0$ for a duration of one time step, with a stimulus value of 1. This was done to assure that each simulated data set could be compared, as even data sets with small values for T would have a covariate that exhibits a peak in the

stimulus function.

The response tensor is simulated from (2.1) with $D = 2$, $\kappa = 0$, and $\sigma^{(0)2} = 1$. Thus, the true coefficient tensor $\mathbf{B}^{(0)}$ is assumed to be sparse and two-dimensional (i.e. $D = 2$). The `specifyregion` function within the `neuRosim` package in R (Welvaert et al., 2011) is employed to simulate the nonzero regions of the true coefficient tensor $\mathbf{B}^{(0)}$. Lengths of each dimension (p_1, p_2) of the tensor coefficient are drawn from a Poisson distribution with shared parameter μ and the nonzero elements assuming value η, which can be thought of as the contrast-to-noise ratio when the observed noise is $\sigma^{(0)2} = 1$. The scenarios were created by constructing a grid over different values for $T \in \{20, 50, 100, 200\}$, $\mu \in \{5, 10, 20, 30\}$, and $\eta = \{0.1, 0.25, 0.5, 0.75, 1, 1.5\}$. For all values of T, the covariate is generated using the `canonicalHRF` function in the `neuRosim` package in R (Welvaert et al., 2011) in which the delay of response relative to onset is $T \times 0.12$, the delay of undershoot is $T \times 0.5$, the dispersion of response was set to 2, the dispersion of undershoot is set to 1, and the scale of undershoot was set to 0.5. This setup was used so that simulations could be run under different values of T without affecting the number of stimulus blocks in the simulated data and without changing the relative pattern of the simulated covariate values.

The model is fitted in each simulation scenario along with the *General Linear Model*, in which the maximum likelihood estimator is found independently for each element in the coefficient tensor $\mathbf{B}$ to highlight the advantages of joint Bayesian modeling with tensor coefficients. In order to mirror the autoregressive error structure in the BTRR models, the `cochrane.orcutt` function in the `orcutt` package in R (Stefano et al., 2018) was used to perform the iterative process necessary to estimate the values of $\mathbf{B}$ and κ. For the Bayesian models, the log-likelihood was examined in order to verify that the Markov chain converged. The

models witness rapid convergence, so that in each model fitting 1,100 draws are taken from the joint posterior distribution, out of which the first 100 draws are discarded as burn-in. Average effective sample sizes shown in Figure 2.2 for the $1,000$ post burn samples calculated using the `coda` package in `R` confirm sufficiently uncorrelated post burn-in samples. The Deviance Information Criterion (DIC) values for each of the shown models can be seen in table 2.1. Note that since there is only one area of activation in each of these datasets, selecting the rank 1 model should produce the best results. This is reflected in the DIC values, as the rank 1 models have the lowest DIC values.

	$\eta = 0.5$	$\eta = 1$	$\eta = 1.5$
Rank 1	14128.27	7089.52	5585.77
Rank 2	14141.91	7090.09	5589.41
Rank 3	14137.61	7093.17	5593.56

Table 2.1: The Deviance Information Criterion corresponding to the estimates in figure 2.3.

Point estimation of $\mathbf{B}$. A comparison of the posterior mean of the elements of $\mathbf{B}$ for different values of R and η when $\mu = 30$ and $T = 20$ can be seen in Figure 2.3. We especially show figures in this case since this case represents higher tensor dimensions and smaller sample size. The posterior mean estimates show the effects of the regularization in the prior, which pulls the posterior mean values corresponding to unimportant cells closer to zero. The true and the estimated activation maps demonstrate excellent performance of BTRR in capturing the true activation pattern under moderate contrast-to-noise ratio η. When contrast-to-noise ratio drops below 1, identifying signal from noise remains a challenging task which causes less accurate identification of activated regions. It should be mentioned that this simulation scenario is well outside the umbrella of theoretical guarantee observed in Theorem 2.3.1 since $s_T \log(p_T)$ is much larger than T, and

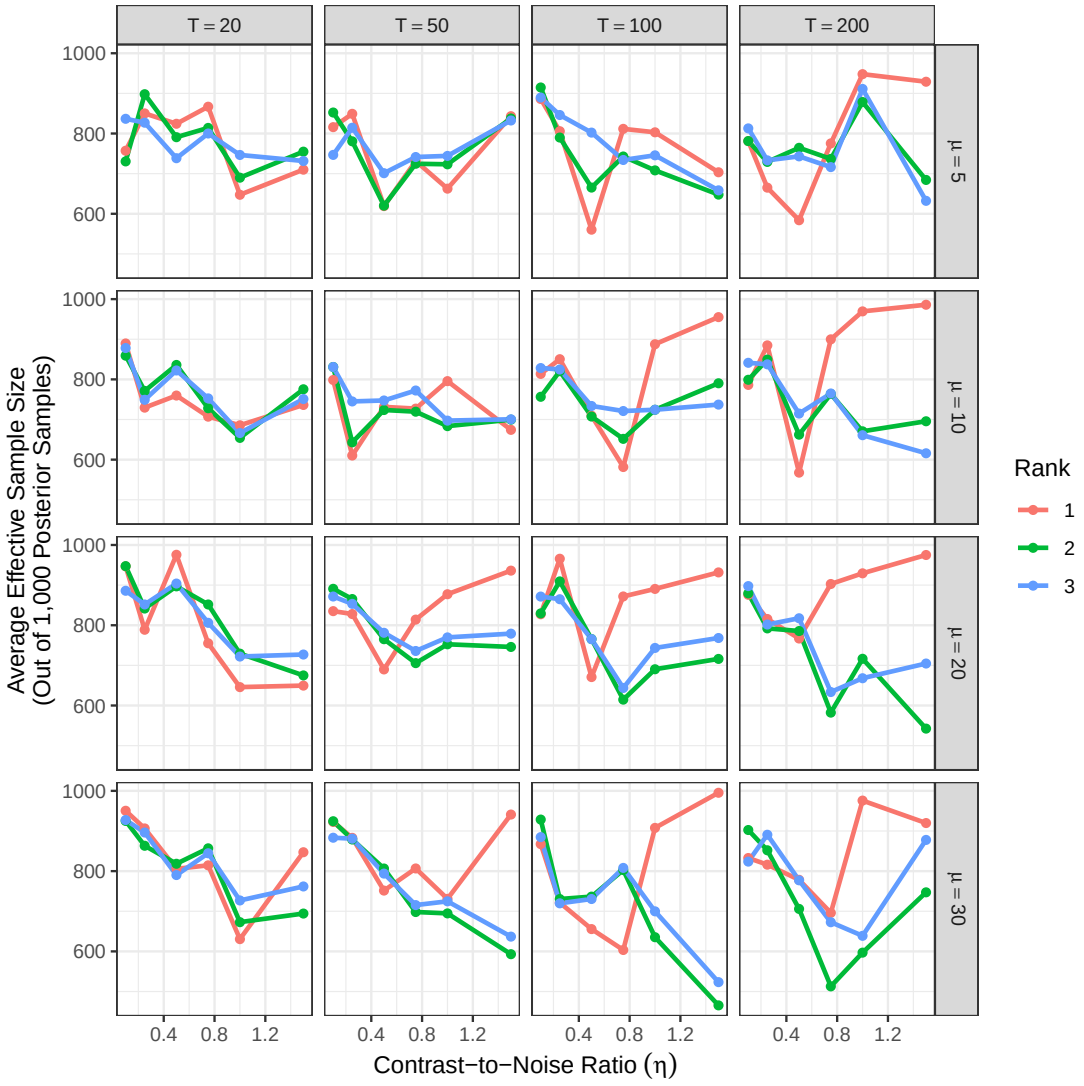

Figure 2.2: The average effective sample size for elements of $\mathbf{B}$ under each of 288 scenarios.

yet the model is able to identify the truly activated regions.

Figure 2.4 shows the root mean squared error of the estimates of $\mathbf{B}$ under different scenarios. In each scenario, BTRR model with ranks $R = (1, 2, 3)$ are tested, and further testing suggests that additional ranks are not required. In a real data application, the final rank used to fit a model can be selected using the deviance information criterion (Gelman et al., 2014). The model at ranks 1, 2, and 3 is

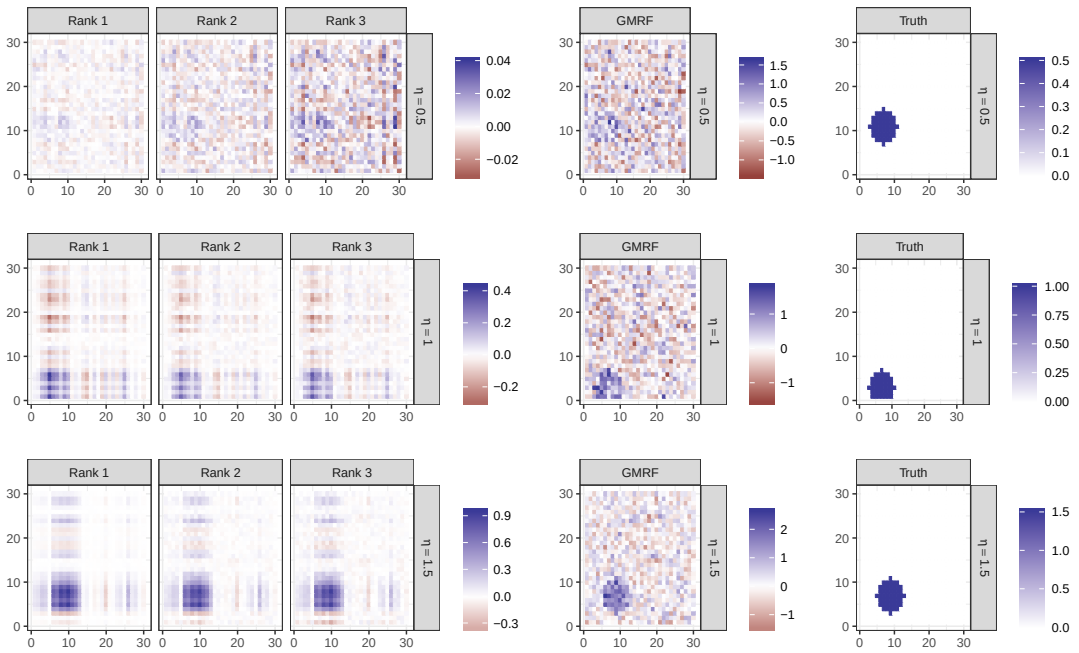

Figure 2.3: Posterior mean and true values for $\mathbf{B}^0$ when $\mu = 30$ and $T = 20$ under different values for R and η. For comparison, the posterior mean estimate from a Gaussian Markov Random Field (GMRF) is also included.

compared to a naive maximum likelihood estimate, which is found by regressing each $\mathbf{Y}_{t,v}$ on $\mathbf{x}_t$ separately for each cell in the experiment.

Based on the results, the model performs well, both for low and high contrast-to-noise ratio. Although the root mean squared error (RMSE) metric seems to be lower for $\eta = 0.5$ compared to $\eta = 1.5$, this does not contradict Figure 2.3. It is worth noting that the shrinkage mechanism pulls every coefficient towards zero, with significant cell coefficients observing less shrinkage than unimportant coefficients. Since for $\eta = 0.5$, even important coefficients are close to zero, all estimated coefficients are close to the truth. On the contrary, for $\eta = 1.5$, shrinkage of important coefficients leads to an increase in RMSE. When RMSE figures are normalized with the true signal strength, the model shows much improved performance for $\eta = 1.5$ than $\eta = 0.5$. Note that the naive MLE does not assume

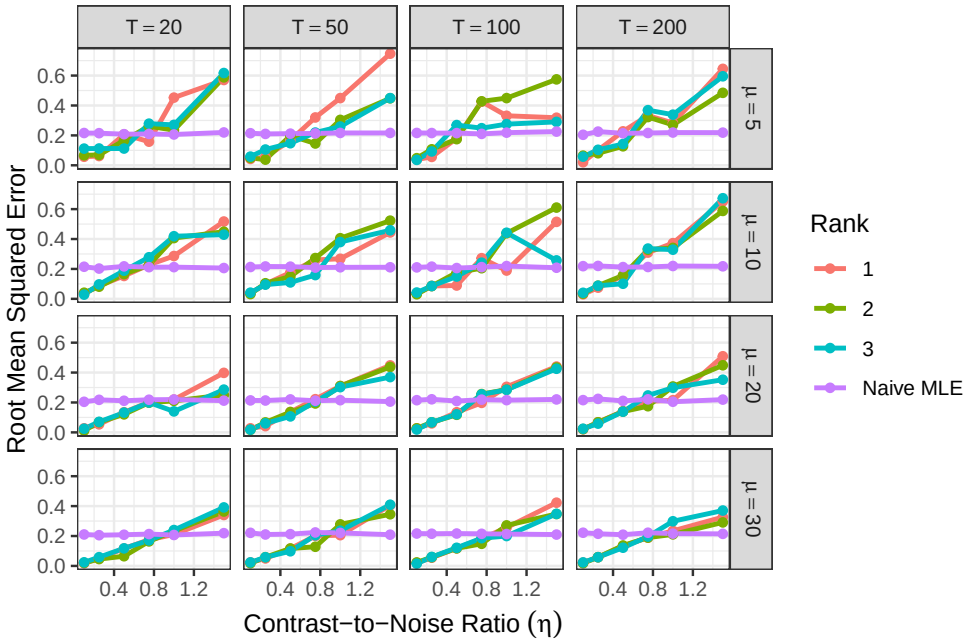

Figure 2.4: Root mean squared error from analyses on simulated data.

sparsity and uses more parameters in almost every case, which is a disadvantage in low-signal sparse regressions. It is used as a comparison to show that the proposed model provides a reasonable point estimate.

A Note on the Use of the Stick-Breaking Structure of $\mathbf{\Phi}_{j,k}$. The stick-breaking structure in the prior of $\mathbf{\Phi}_{j,k}$ allows for a flexible modulation of the effects of additional ranks on the posterior densities of the elements within $\mathbf{B}_k$ at a very small cost in computational efficiency. In a test with a simulated dataset with a single covariate, $\mathbf{Y}_t \in \mathbb{R}^{30\times 30}$ and $T = 100$ with model rank 3, the cost of the posterior updates for all parameters associated with the stick breaking structure accounted for just over 2% of the total computation time.

Comparison to Gaussian Markov Random Field model. Tensor regression data could also be modeled with spatial dependence through a Gaussian Markov Random Field (GMRF) (Zhang et al., 2015; Gössl et al., 2001; Quirós et al., 2010). For comparison, a model was created with a GMRF prior on the vectorized elements

of $\mathbf{B}$. That is,

$$\begin{aligned}
\text{vec}\mathbf{B} &\sim \text{N}(\mathbf{0}, (\lambda\mathbf{Q})^{-1}), \\
\lambda &\sim \text{Gamma}(a_\lambda, b_\lambda), \\
\mathbf{Q} &= \begin{cases} n_v, & v = \ell \\ -1, & v \sim \ell \\ 0, & \text{otherwise} \end{cases},
\end{aligned}$$

where n_v is the number of neighbors of element v and $v \sim \ell$ denotes that elements v and ℓ are neighbors (Zhang et al., 2015). In the case of these simulations, a_λ was set to 1 and b_λ was set equal to 0.001. This was done to match the noninformative prior in the BTRR model. The results in figure 2.3 show that while the GMRF model can identify the region of activation, it does so with a higher variance in the case when $\eta \in \{1, 1.5\}$, resulting in less precise inference. Thus, the BTRR model is able to impose more sparsity and more precisely identify contiguous areas within $\mathbf{B}$ that have nonzero values.

Parametric uncertainty of $\mathbf{B}$. To assess uncertainty quantification of $\mathbf{B}$ from BTRR, we focus on coverage and length of 95% credible intervals (CI) of cells of $\mathbf{B}$, shown in Figures 2.6 and 2.7 respectively. Given that in almost all the scenarios, the coverage of the 95% credible intervals is close to nominal, attention turns to the length of the 95% credible intervals. Two visible patterns emerge from the figures. First, the 95% credible intervals shrink as T increases, since the posterior variance lowers with increased observed data. Secondly, the credible intervals are wider for higher contrast-to-noise ratio, which can be attributed to the fact that estimating a few high signals with lots of zero coefficients involves more uncertainties. Finally,

there is a drop in coverage for the 95% posterior credible intervals as the contrast-to-noise ratio increases, especially for small tensor dimensions. This is due to the fact that very low values for η mean that the true values for the elements in $\mathbf{B}$ are very close to zero. Therefore, in conjunction with the regularization of the parameter estimates stemming from the prior structure, many of the true values are close enough to zero that they are all captured by credible intervals centered at or near zero.

Simulations under an autoregressive error distribution. While some applications of the BTRR model may have an error structure that is independent through time, such as the application in section 2.5, such an assumption may not be tractable in other situations. In order to show that the model still performs well in the face of a stationary temporal error structure, tests were performed on a data set with an error structure that is autoregressive with order 1. Data were simulated assuming $T = 100$, $\mu = 30$, $\eta = 1$, and $\kappa = 0.5$. Plots showing the estimated values for $\mathbf{B}$ and the posterior density estimates for the autoregression coefficient from 1,000 samples from the posterior distribution can be seen in figure 2.5.

All scenarios conclusively establish the strength of BTRR as a principled Bayesian approach that accurately detects brain activation with proper characterization of uncertainties. It is particularly appealing to observe BTRR outperforming MLE estimates in smaller contrast-to-noise ratios reminiscent of real fMRI data. We have also replaced naive MLE by a few penalized optimizers (not shown here) and found BTRR continuing to be the clear winner in sparse and low signal scenarios. Application of the model to real data is explored in the following section.

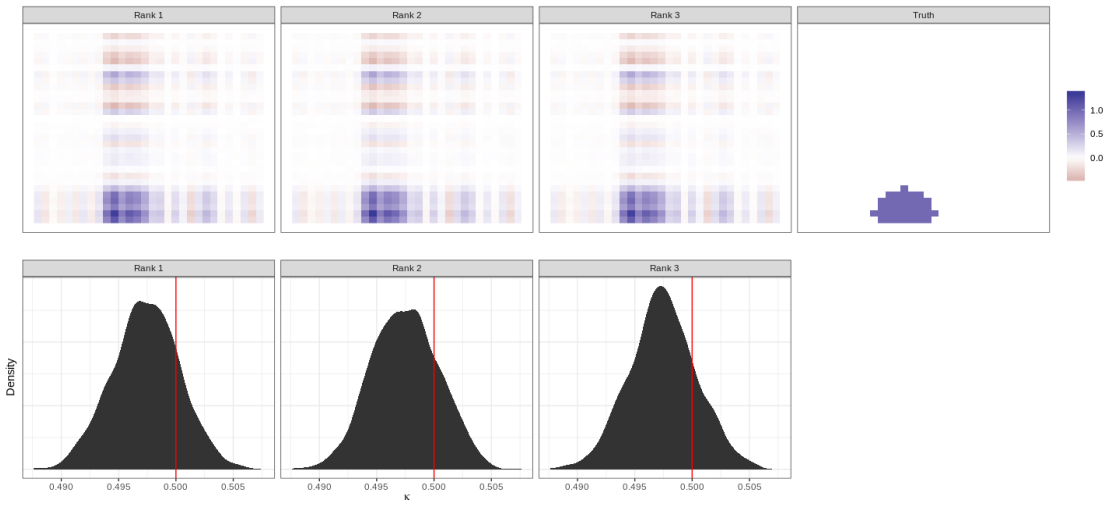

Figure 2.5: Plots of the posterior means of $\mathbf{B}$ next to the true value (top) and the posterior densities of the autoregression coefficient (bottom). The true value for the autoregression coefficient is indicated with a red line.

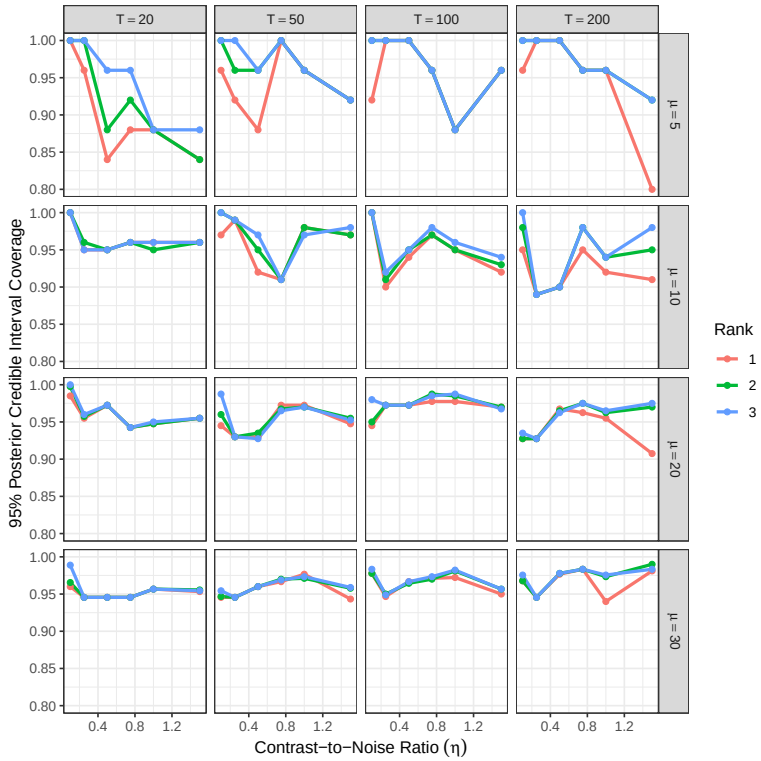

Figure 2.6: The average coverage of the 95% posterior credible intervals for the posterior draws for the elements of $\mathbf{B}$ under varying conditions.

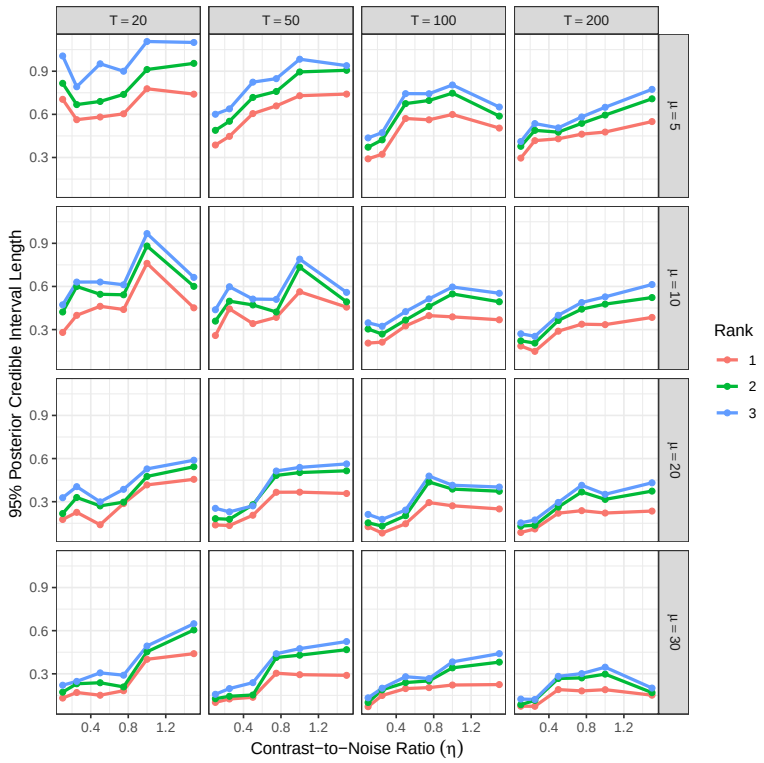

Figure 2.7: The average length of the 95% posterior credible intervals for the posterior draws for the elements of **B** under varying conditions.

2.5 Application to Balloon Analog Risk Taking Data

Neuroscientists at the University of California Los Angeles conducted an experiment intended to make inference about the regions of the brain that are involved in the process of evaluating risk (Schonberg et al., 2012). Sixteen young adults (average age of 23.56 years) were subjects in an experiment with the following design. Each subject entered an fMRI machine with a computer display and a controller with two buttons. On the screen, the image of a balloon would be shown, along with a payout amount, starting with a value of \$0.25. The buttons on the controller allowed the subject to either inflate the balloon or take the

payout. If the subject inflated the balloon, and the balloon did not explode, the payout amount increased by \$0.25. If the subject inflated the balloon, and the balloon exploded, no payout was received, the payout value was reset to \$0.25, and a new balloon was displayed. Balloons were assigned a number of pumps at which the balloon would explode from a discrete uniform distribution with a lower bound of 1, and an upper bound of 8, 12, or 16, depending on whether the balloon was red, green, or blue, respectively. A grey "control" balloon, offering no payout and an upper bound of 12 pumps before exploding, was also part of the trial to record a riskless scenario. Each subject participated in three runs. Each run consisted of either 10 minutes, or 48 balloons exploding, whichever came first.

The proposed method is designed for single-subject data, and thus data from a single run for one subject were analyzed. Before analysis, the data were preprocessed to correct for motion and other nuisance variables, and to map the subject's brain into a standard space using FSL (Smith et al., 2004). The resulting images were then sliced into a two-dimensional cross-section and separated into 9 different regions of interest based on the MNI atlas from the Montreal Neurological Institute, which is distributed with the FSL software. Further details on the fMRI image preprocessing can be found in the appendix. This separation was done in order to speed up computation time by parallelizing the analysis of different regions, and also to allow for different values of R to be selected depending on the needs of a specific region. As regions of interest are not box-shaped in nature, the smallest box-shaped region containing the region of interest was taken as the response tensor. Parts of the box-shaped region that were not a part of the region of interest were all assigned the value 0, which does not impact the inference on the coefficients, as the multiway stick-breaking prior is able to assign the coefficient tensor values around zero. Table 2.2 records these Regions of Interest (ROI)

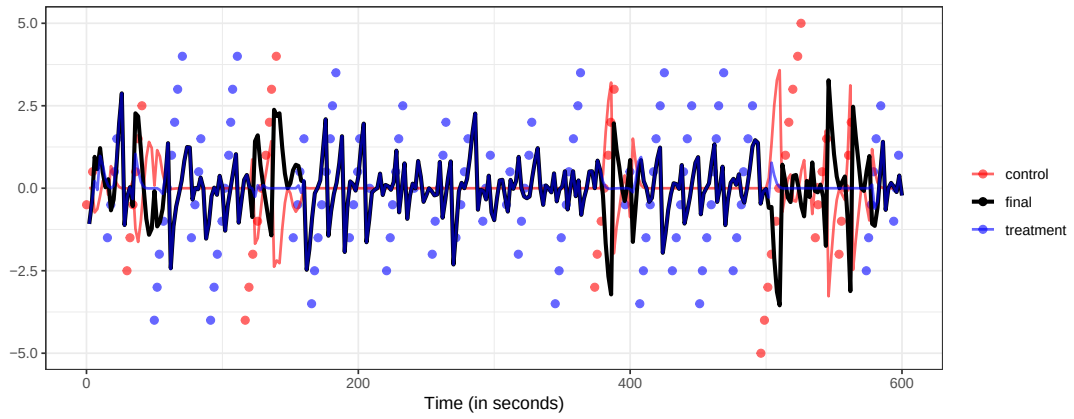

Figure 2.8: The raw values of the demeaned number of pumps (points), their convolution with the double-gamma haemodynamic response function (light lines), and the final covariate resulting from their difference (heavy black line) for the subject analyzed.

with the number of voxels in each. Finally, the first difference was taken for each voxel-level time series in order to account for scanner drift, a phenomenon caused by the fMRI scanner's magnetic field drifting over time (Huettel et al., 2004). Our preliminary investigation of the resulting data confirms that fitting the ROIs separately performs better than a single analysis of the whole brain due to differing observed variance in different regions.

To measure the level of risk to a subject at a given time, we follow the procedure used in Schonberg et al. (2012), the study from which these data were found. First, we measure the centered number of pumps that an individual gives a "treatment" balloon before they "cash-out" or the balloon explodes. It is assumed that the higher the number of pumps becomes, the subject experiences a higher perceived risk. This value is then convolved with the double-gamma haemodynamic response function, which takes into account the physiological lag between stimulus and response, and smooths the stepwise function for the centered number of pumps. An illustration of this calculation can be seen in Figure 2.8.

Finally, the centered, convolved number of pumps on the control balloon is

subtracted from the treatment series to provide a basis for comparison. The multiway stick-breaking shrinkage model as defined in 2.1 is applied to this data under ranks 1, 2, and 3 in each region, and each Markov chain is run for 1,100 iterations. As a note, no additional parameters were added to the model to correct for slice timing and other potential nuisance signals. Such an extension of the model may be explored in future work. The first 100 iterations are discarded as burn-in after checking the stationarity of the log-likelihood. The point estimates for activation coefficient for different ranks can be seen in Figure 2.9. As the brain images are split into nine regions of interest and BTRR fitted separately for these regions, the final estimate for activation coefficient $\mathbf{B}$ is obtained by using the sequential 2-means post-processing algorithm proposed by Li and Pati (2017) in each ROI, which is discussed further in chapter 3. The rank R used for estimation in each ROI is chosen using the Deviance Information Criterion (DIC) (Gelman et al., 2014). The final estimate obtained in this fashion is presented in Figure 2.10.

The same general linear model maximum likelihood estimate described in section 2.4 is also added for comparison. The general linear model has estimates that are somewhat larger than the Bayesian sparse tensor response regression estimates. This is likely due to the shrinkage prior favoring smaller values. However, the final point estimates from the general linear model and the BTRR models show generally coincident voxel-level effects, with the BTRR models showing fewer active voxels due to the regularization that they impose and the sequential 2-means variable selection method. This likely reduces the false positive rate of activation detection, which improves inference on the experiment overall.

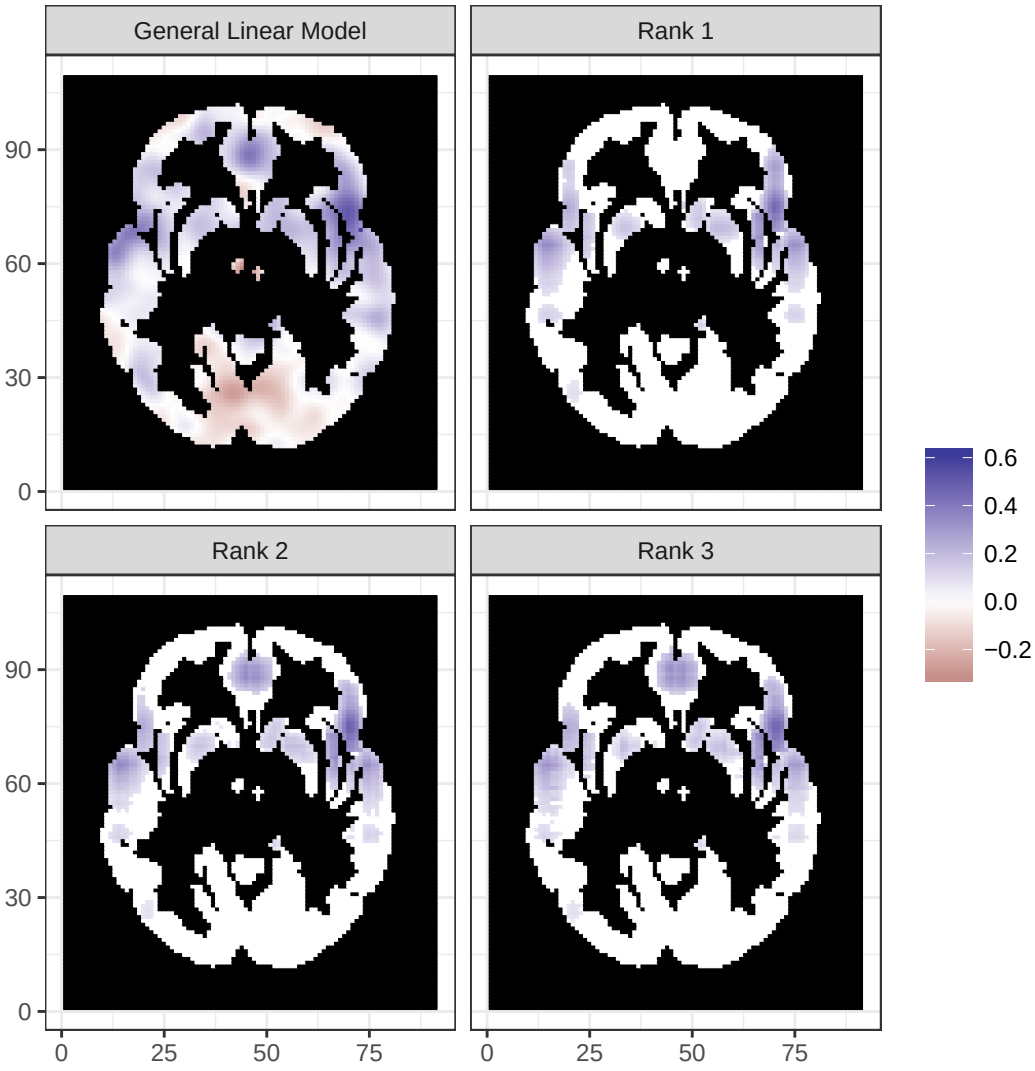

Figure 2.9: A comparison of estimates of $\mathbf{B}$ under the general linear model maximum likelihood estimate, and the Bayesian Sparse Tensor Response Regression models with ranks 1, 2, and 3.

2.6 Conclusion

This chapter proposes a Bayesian framework to regress a tensor valued response on scalar covariates. Adopting the rank-R PARAFAC decomposition for the tensor coefficient, the proposed model is able to reduce the number of free

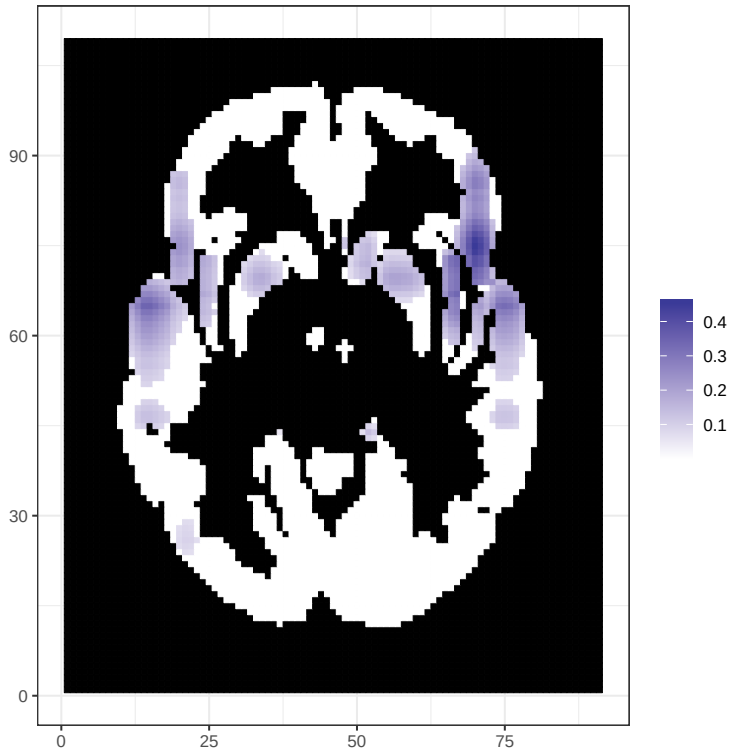

Figure 2.10: The final estimate of the effect of increased perceived risk on the relative levels of oxygen in different regions of the brain after selecting R for each region using the DIC.

parameters. We employ a novel multiway stick-breaking shrinkage prior distribution on the tensor coefficient to be able to identify significantly nonzero cell coefficients. New results on posterior consistency have been developed to show convergence in L_2 sense of the tensor coefficient to the true tensor as data size increases.

As an illustrative example, the present chapter focuses on analysis of fMRI data to detect voxels of the brain which exhibit neuronal activity in response to stimuli, while simultaneously inferring on the association of spatially remote groups of voxels with similar characteristics. Analysis of simulated fMRI time series and real fMRI data demonstrates excellent performance of BTRR in identifying the

Region	p_1	p_2	Chosen Rank
Cerebellum	8	8	Rank 1
Putamen	34	19	Rank 1
Parietal Lobe	22	7	Rank 1
Caudate	19	9	Rank 1
Frontal Lobe	57	35	Rank 1
Insula	45	23	Rank 1
Occipital Lobe	56	31	Rank 1
Temporal Lobe	72	40	Rank 1
Thalamus	8	6	Rank 3

Table 2.2: The different values for R selected by the deviance information criterion (DIC), along with the dimensions associated with the response tensors in each region.

regions of activation with required uncertainties. Additionally, BTRR is able to achieve remarkable parsimony, even as a Bayesian model. This facilitates its usage in presence of images with a fine resolution.

The core idea of the proposal is to recognize the importance of retaining the tensor structure of the image response during the entire statistical analysis for studies including brain activation. An immediate extension to the proposed model would be meant to investigate both voxel-level activation and ROI-level connectivity from multi-subject fMRI data. An additional extension may be to explore an expansion of the error characterization to include spatial dependence in addition to temporal dependence.

Chapter 3

Joint Bayesian Estimation of Voxel Activation and Interregional Connectivity in fMRI

Next, we extend the methods from chapter 2 to a scenario with multiple tensor responses for multiple subjects, which increases the complexity of the model. Now we are not only interested in the tensor-valued coefficients, but also the correlations between the means of the response tensors. This increases the complexity of the model, but also produces a rich inference that will be used to analyze all of the subjects from the fMRI experiment in chapter 2.

3.1 Introduction

Expected activation patterns in task-based fMRI experiments include local spatial dependence in the sense that voxels located next to each other tend to be

jointly activated, as well as non-local dependencies in which groups of voxels in distant regions of the brain are activated by a given thought process.

In addition to determining brain activation linked with a specific cognitive or sensorimotor function, neuroscientists are often interested in the way in which different spatially-adjacent groups of voxels, referred to as Regions of Interest (ROIs) in the brain work together to process information. These types of relationships between ROIs are collectively referred to as *functional connectivity* (Hutchison et al., 2013). The major contribution of this article is the proposal of a Bayesian modeling framework that simultaneously detects voxel-level activation and connectivity between different ROIs with precise characterization of uncertainty for multi-subject fMRI data. These methods, applied to task-based fMRI experiments, lead to meaningful inferences about the functions of the human brain in such experimental settings.

Simultaneous analysis of multi-subject 3D fMRI scans is a challenging problem due to the sheer amount of data. Our modeling framework addresses this issue by using a mixed-effects tensor response regression analysis in which low-rank tensor decompositions are combined with a multiway stick-breaking shrinkage prior to achieve parsimony in the estimation of voxel-level activation. Such framework provides a powerful and computationally-feasible setting for inferring activation and connectivity in multi-subject task-related fMRI studies.

Before describing our proposed modeling approach, we present an overview of currently available methods that separately infer activation at the voxel level and connectivity at the region-specific level in Sections 3.1.1 and 3.1.2 respectively. Section 3.1.3 then presents a review of models that jointly infer activation and connectivity. Due to the space constraint, we mainly focus on describing approaches that would be natural competitors to our proposed approach either

because they share similarities in at least one of the modeling components (e.g., tensor-based approaches), or they use the same inferential paradigm (Bayesian approaches).

3.1.1 Activation Models

Several approaches have been proposed for the analysis of brain activation. Single-subject frameworks in particular have a rich background for modeling activation. The simplest of them, known as the General Linear Model (GLM), fits a regression model at each voxel with the observed voxel-specific BOLD response regressed on activation related predictors and identifies if the response is significantly associated with the predictors in that voxel (Friston et al., 1995; Penny et al., 2011), after accounting for multiple testing corrections. Many implementations of the GLM include clusterwise corrections to account for spatial association using some variant of independent components analysis. However, work by Eklund et al. (2016) suggests that many of these methods inflate false-positive rates. The Banjamini-Hochberg correction (Benjamini and Hochberg, 1995), which works by setting the false discovery rate, is a possible solution, but remains incomplete, as it does not take spatial information into account. Another idea, which addresses the sparse nature of fMRI activation (Olshausen and Field, 2004), assigns a spike-and-slab prior on the regression coefficients (Brown et al., 1998; Yu et al., 2018) across all voxels. These priors take the form

$$\beta_v \sim \gamma_v \text{Normal}(0, v_1) + (1 - \gamma_v)\text{Normal}(0, v_0), \tag{3.1}$$

with β_v the activation parameter at voxel v. In this setting, v_0 is relatively small, v_1 is relatively large. γ_v is a zero-one random variable such that $Pr(\gamma_v = 1) = \pi_v$, with π_v being the probability that β_v comes from the normal distribution with

larger variance. While these methods tend to have robust sensitivity and specificity, they do not take spatial information into account. In fact, brain activation is a local process in the sense that spatially contiguous voxels generally tend to be activated together while performing a task. Spike-and-slab prior distribution is unable to incorporate such structural information a priori.

Various approaches also account for spatial association in the neighboring voxels by inducing dependence among voxel-specific regression coefficients using Markov random fields (Zhang et al., 2014b; Kalus et al., 2014; Smith and Fahrmeir, 2007; Lee et al., 2014). In these approaches, the activation coefficients are represented as having a multivariate normal prior with precision $\mathbf{\Sigma}^{-1}$. Within $\mathbf{\Sigma}^{-1}$ the values on the diagonal are the number of neighbors that voxel v has. The off-diagonal elements σ_{vk} are set equal to 1 if voxel v and voxel k are considered neighbors, and 0 otherwise. Models with this type of Markov random fields components usually capture, at least partially, the underlying spatial structure in fMRI data but tend to be computationally expensive. More complex Markov random field structures may be proposed, but almost certainly at the cost of further decreased computational efficiency. In contrast, our proposed tensor mixed effect model improves inference by encouraging localized activation without explicitly employing spatial dependence.

Other sophisticated approaches include spatially varying coefficient (SVC) models which employ spatial basis functions to model activation-related coefficients (Flandin and Penny, 2007; Zhu et al., 2014). Besides being computationally expensive, such models are sensitive to the selection of the basis functions, and require specific knowledge to appropriately calibrate them.

More recently, tensor-based approaches have been proposed and applied to analyze large-dimensional brain imaging data. Zhou et al. (2013) considers a frame-

work that linearly models a given clinical outcome as a response and uses brain images as tensor covariates. Estimation in this setting is achieved via maximum likelihood and regularized tensor regression tools based on tensor decomposition are also used to determine which regions of the brain are associated with the clinical response. Tensor decomposition methods are covered in further detail in Section 3.2. Guhaniyogi et al. (2017) proposes a Bayesian tensor regression method in presence of a scalar response and a tensor predictor with shrinkage priors to identify cells in the tensor predictor significantly related to the scalar response. Tensor-based frequentist and Bayesian approaches for joint modeling of the BOLD response across all voxels in the form of a tensor have also been developed, as in the previous chapter (e.g., Li and Zhang, 2017). However, these approaches have not been developed for multi-subject studies and do not allow us to infer connectivity between brain regions.

3.1.2 Multi-subject and Connectivity Approaches

In order to overcome the computational challenge of having voxel-level data analyzed for multiple subjects, early approaches combined information across voxels, either using the general linear model (GLM) parameter estimates or residual variance. Two-stage methods, such as those by Bowman et al. (2008) and Sanyal and Ferreira (2012), fit subject-specific GLMs, and then use regularization on the parameter estimates to determine activation. Mixture models and non-parametric Bayesian models (Zhang et al., 2016; Xu et al., 2009) have also been proposed to analyze inter- and intra-subject variability, though they incur heavy computational cost for exact results via Markov Chain Monte Carlo (MCMC).

While there is considerable literature on activation-only models, literature on Bayesian functional connectivity offers a much smaller number of distinct ap-

proaches. Most of the available models consider connectivity across clusters of voxels that display similar activity patterns, or develop connectivity measures across predetermined ROIs. To this end, Patel et al. (2006a) and Patel et al. (2006b) discretized the fMRI time series between regions based on whether they had elevated activity according to a threshold, and then compared joint and marginal probabilities of elevated activity. Bowman et al. (2008) modeled similarity within and between ROIs based on estimates of elements of the covariance in a two-stage model. A Dirichlet process prior was used by Zhang et al. (2014b) to cluster remote voxels together, asserting that the clustering inferred an inherent connectivity. Zhang et al. (2014a) went on to propose a dynamic functional connectivity model, estimating connective phases and temporal transitions between them.

3.1.3 Joint Estimation of Activation and Connectivity

As mentioned above, there are several Bayesian modeling frameworks for assessing activation or connectivity separately, however, models incorporating both of them jointly in multi-subject fMRI studies with voxel-level data are comparatively rare in the literature. In the recent past, Kook et al. (2017) proposed an approach in which a Dirichlet process (DP) mixture model is used to classify voxels as active or inactive via discrete wavelet transformations. The clustering of the voxels through time via the mixture components is then used to derive a measure of inter-voxel connectivity within- and between-subjects. While such model succinctly captures activation and connectivity, the use of a Dirichlet process hinders computational efficiency. Variational Bayes methods were used to speed-up computation, providing results in a fraction of the time that a full Markov chain Monte Carlo simulation would require, however these variational approaches lead to approximate rather than exact posterior inference. In addition, the model of Kook

et al. (2017) analyzes voxel-level connectivity, rather than region-of-interest connectivity. Connectivity-informed activation detection was proposed by Ng et al. (2011) using a two-step classical modeling process. First, a graphical LASSO or oracle-approximating shrinkage is used to cluster voxels into groups, and then cluster information is incorporated into the estimation for activation effects. This model differs from our proposal in that it does not provide robust uncertainty quantification, and regions can not be defined *a priori*. In addition, activation is detected through an average over parcellated regions, and as such, voxel-level activation inference is not possible.

The Potts model (Potts, 1952) could be considered as an alternative to either the activation or connectivity model components. The model operates by defining a fixed number of states, as well as a neighborhood structure for the elements within the tensor coefficients or posterior precision between regions of interest. When the number of states is equal to 2, the model is equivalent to the Ising model (Ising, 1925), in which model parameters would simply be classified as zero or nonzero. Using the tensor decomposition structure in conjunction with the generalized double-Pareto prior allows for judgements not just about activation, but also about effect size in a continuous setting. The Potts model on the connectivity would likely be problematic through the bias introduced by defining neighborhoods and the number of states within the precision matrix.

Our article proposes a multi-subject Bayesian tensor mixed effect model that provides exact posterior inference on voxel-level activation using the inherent spatial structure of fMRI and region-of-interest-level connectivity measures in a single model, while imposing shrinkage on both measures. Modeling activation and connectivity in a single step reduces model bias over a two-step modeling process by taking into account any interaction between the activation and connectivity

components. In addition, our approach does not require a multiplicity correction method when detecting activation and/or connectivity. To elaborate further, the model envisions the BOLD response over all voxels together for a subject at any time as a tensor and regresses this tensor object on the activation related predictors. A tensor decomposition representation is then used in conjunction with a novel prior structure to make the model more parsimonious while simultaneously capturing the underlying spatial structure of the data in an efficient manner. This is one of the key features of using a tensor representation, i.e., it enables the use of a tensor decomposition structure, which essentially treats the principle axes of the coordinate system as principal components, inherently preserving localized spatial dependence while reducing the number of model parameters. This is an advantage with respect to approaches such as those in Kook et al. (2017) which explicitly model local spatial correlation among voxels using spatially dependent prior distributions on voxel specific activation coefficients, and hence becoming computationally prohibitive with a large number of voxels, particularly in large multi-subject studies. In addition, our proposed model incorporates subject-ROI-specific random effects with a Gaussian graphical prior, imposing regularization on the precision matrix of the effects between regions (Wang et al., 2014). Both, the activation and connectivity parameters are then classified into zero- and nonzero-effect sizes using the sequential 2-means method proposed by Li and Pati (2017). As a result, the model produces accurate measures of voxel-wise activation and inter-regional connectivity with interpretable effect sizes without the need for fine-tuning hyperparameters, basis functions or multiplicity corrections. In addition, the model is computationally tractable to provide samples from the exact posterior distribution for 2-D slices or 3-D volumes of brain images, as well as higher-order tensor images.

The upcoming sections proceed as follows. The model, including the prior structure, is set forth in the Section 2. This is followed by a section on posterior inference (Section 3) and a section on simulated studies to empirically validate the model (Section 4). Sensitivity to hyperparameter specification and comparisons with other models is also shown in Section 4. Section 5 then describes the multi-subject fMRI data from the balloon-analog risk-taking experiment in detail, as well as the results obtained from applying the proposed methodology to these data. Finally, Section 6 summarizes our findings and provides a description of some future extensions.

3.2 Methodology

This section begins by detailing the model framework and prior structures for parameters estimating the activation and connectivity. A recommended setting for the hyperparameter values is then outlined in order to provide reliable inference.

3.2.1 Model framework and prior structure

We assume that the whole tensor structure of the fMRI is partitioned into G distinct brain regions, and that we are interested in effectively measuring brain connectivity between these regions. Importantly, the proposed model does not assume that the voxel-level activation and the functional connectivity between predefined regions of interest are independent. As an aside, note that setting $G = 1$ reduces the model to a simpler tensor response regression model, which is explored in a single-subject context in the previous chapter. Our proposal extends that model to a multi-subject tensor response regression setting. In addition, the proposed model also takes inter-regional connectivity into account.

Let $\boldsymbol{Y}_{i,g,t}$ be the tensor of observed BOLD response data in brain region g for the ith subject at the tth time point. $\boldsymbol{Y}_{i,g,t}$ is observed in the form of a tensor with dimensions $p_{1,g} \times \cdots \times p_{D,g}$. In the context of fMRI data analysis, the tensor dimension for a fixed time, subject, and region, denoted as D, is two or three, depending on whether a single slice or regional volume is analyzed. To simultaneously measure activation due to stimulus at voxels in the gth brain region and connectivity among G brain regions, we employ an additive mixed effect model with tensor-valued BOLD response and activation-related predictor $x_{i,t} \in \mathbb{R}$,

$$\mathbf{Y}_{i,g,t} = \mathbf{B}_g x_{i,t} + d_{i,g} + \mathbf{E}_{i,g,t}, \tag{3.2}$$

for subject $i = 1, \ldots, n$, in region of interest $g = 1, \ldots, G$, and time $t = 1, \ldots, T$. Elements in the error tensor $\mathbf{E}_{i,g,t}$ are assumed to be independent and identically distributed following a normal distribution with mean 0 and shared variance σ_y^2, though our framework can be extended to incorporate temporally correlated errors. However, in testing with both simulated and task-based fMRI data, this does not appear to have a large effect on the model inference.

The model in (3.2) has 3 components, namely, an activation component, a connectivity component and an error component. The tensor coefficient $\mathbf{B}_g \in \mathbb{R}^{p_{1,g} \times \cdots \times p_{D,g}}$ is used to infer the strength of the association between $x_{i,t}$ and each voxel in $\mathbf{Y}_{i,g,t}$. In particular, $B_g[i_1, ..., i_D] = 0$ implies that the $(i_1, ..., i_D)$th voxel in the gth ROI is *not activated* by the stimulus. In fact, the activation pattern is typically sparse and localized with only a few nonzero elements in $\boldsymbol{B}_g$ (Olshausen and Field, 2004). $d_{i,g} \in \mathbb{R}$ are region- and subject-specific random effects which are jointly modeled to borrow information across ROIs. This model views connectedness through the elements of the precision matrix corresponding to different, pre-specified regions of interest, rather than between individual voxels. In

this way, the detected relationship between regions is not directly determined by the detected activation of a voxel. In the present context, the conditional distributions $(d_{i,g}, d_{i,g'})|\{d_{i,g''} : g'' \neq g, g'\}$ are investigated to assess the strength of connectivity between a pair of regions. As part of the model development, we impose prior distributions that favor conditional independence between most pairs $d_{i,g}$ and $d_{i,g'}$, effectively favoring connectivity only among a few pairs of regions. Choosing a different number of regions will obviously change the network of regions that inference can be applied to and also the inference in the model parameters. However, since the partial correlation is used to measure connectivity, the connectivity measured between two given regions should not be greatly affected by adding or removing other regions from the model.

As mentioned above, the coefficient tensor $\mathbf{B}_g \in \mathbb{R}^{p_{1,g} \times \cdots \times p_{D,g}}$ in equation (3.2) characterizes a sparse relationship between the tensor response and the time-varying covariate $x_{i,t}$ in region g. In order to achieve parsimony in the number of estimated parameters, $\mathbf{B}_g$ is assumed to have a rank R parallel factorization (PARAFAC) decomposition, which takes the form:

$$\mathbf{B}_g = \sum_{r=1}^{R} \boldsymbol{\beta}_{g,1,r} \circ \cdots \circ \boldsymbol{\beta}_{g,D,r}, \tag{3.3}$$

with tensor margin effects $\boldsymbol{\beta}_{g,1,r}, .., \boldsymbol{\beta}_{g,D,r}$. The PARAFAC tensor decomposition dramatically reduces the number of parameters in $\boldsymbol{B}_g$ from $\prod_{j=1}^{D} p_{j,g}$ to $R \sum_{j=1}^{D} p_{j,g}$, where the level of parameter reduction depends on R. Note that a smaller value of R leads to more parsimony and computational gain, perhaps at the cost of inferential accuracy. By contrast, a choice of even moderately large R entails higher computation cost. Using R as a model parameter often increases computation cost and is deemed unnecessary (Guhaniyogi et al., 2017). In view of the earlier literature, this article proposes fitting the model with various choices of R and

chooses the one that yields the lowest Deviance Information Criterion (DIC) (Gelman et al., 2014). More discussion on the choice of R is provided in the Simulation Studies section.

A critical question remains how to devise a prior distribution on the low-rank decomposition (3.3) to facilitate identifying geometric sub-regions in the tensor response which share an association with the predictor. Additionally, the model intends to build joint priors on region specific random effects $d_{i,g}$s to assess connectivity patterns. The next two subsections propose careful elicitation of the prior distributions on $\boldsymbol{B}_g$ and $d_{i,g}$ to achieve our stated goals.

3.2.2 Multiway stick breaking shrinkage prior on B_g to assess activation

Although the spike-and-slab priors for selective predictor inclusion (George and McCulloch, 1993; Ishwaran et al., 2005) possess attractive theoretical properties and an easy interpretation, they often lose their appeal due to their inability to explore a large parameter space. As a computationally-convenient alternative, an impressive variety of shrinkage priors (Carvalho et al., 2010; Armagan et al., 2013a) in the context of ordinary Bayesian high dimensional regression have been developed.

Shrinkage architecture relies on shrinking coefficients corresponding to unimportant predictors, while maintaining accurate estimation with uncertainty for important predictor coefficients. The existing shrinkage prior literature serves as a basis to the development of shrinkage priors on the tensor coefficients. However, constructing such a prior on $\boldsymbol{B}_g$ presents additional challenges. To elaborate on it, notice that proposing a prior on a low-rank PARAFAC decomposition of $\boldsymbol{B}_g$ is equivalent to specifying priors over tensor margins $\boldsymbol{\beta}_{g,j,r}$. Since every cell

coefficient in $\boldsymbol{B}_g$ is a nonlinear function of the tensor margins, careful construction of shrinkage priors on $\boldsymbol{\beta}_{g,j,r}$s is important to impose desirable tail behavior of $B_g[i_1, ..., i_D]$ parameters. To this end, this article employs *a multiway stick-breaking* shrinkage prior on $\boldsymbol{B}_g$ to ensure desirable tail behavior. More specifically, the following shrinkage prior is proposed on the tensor margins

$$\boldsymbol{\beta}_{g,j,r} \sim \mathrm{N}\left(\mathbf{0}, \phi_{g,r}\tau_g \mathbf{W}_{g,j,r}\right), \ \mathbf{W}_{g,j,r} = \mathrm{diag}\left(w_{g,j,r,1}, \ldots, w_{g,j,r,p_j}\right),$$

where

$$w_{g,j,r,\ell} \sim \mathrm{Exp}\left(\frac{\lambda_{g,j,r}^2}{2}\right), \ \lambda_{g,j,r} \overset{iid}{\sim} \mathrm{Gamma}(a_\lambda, b_\lambda),$$

for $j = 1, ..., D$ and $g = 1, ..., G$. Flexibility in modeling tensor margins are accommodated by introducing $\boldsymbol{W}_{g,j,r}$s. In fact, integrating out $\boldsymbol{W}_{g,j,r}$ and $\lambda_{g,j,r}$ yields a generalized double Pareto shrinkage prior for the elements of $\boldsymbol{\beta}_{g,j,r}$ conditional on $\phi_{g,r}$. The proposed prior defines a set of rank specific scale parameters $\phi_{g,r}$ using a stick breaking construction of the form $\phi_{g,r} = \xi_{g,r} \prod_{l=1}^{r-1}(1 - \xi_{g,l})$, $r = 1, ..., R-1$, and $\phi_{g,R} = 1 - \sum_{r=1}^{R-1} \phi_{g,r} = \prod_{l=1}^{R-1}(1 - \xi_{g,l})$ that achieves efficient shrinkage across ranks, where $\xi_{g,r} \overset{iid}{\sim} Beta(1, \alpha_g)$. Finally, the global scale parameters are modeled as $\tau_1, ..., \tau_G \overset{iid}{\sim} \mathrm{Gamma}(a_\tau, b_\tau)$.

Without constraints on the values for $\phi_{g,r} \in \boldsymbol{\Phi}_g$, where $r = 1, \ldots, R$, identifiability issues arise in the posterior sampling for the variance terms for $\boldsymbol{\beta}_{g,j,r} \in \boldsymbol{B}_g$. In order to address this issue, a stick-breaking structure is imposed on $\phi_{g,r}$'s, as described above. In effect, this prevents $\phi_{g,r}$'s from switching labels across ranks in which the variance of $\boldsymbol{\beta}_{g,j,r}$ may be close together. The result of this constraint is a more stable MCMC for the posterior draws of $\boldsymbol{\beta}_{g,j,r}$. The tuning parameter α_g in the stick-breaking construction determines which tensor rank R is favored by

data. In particular, $\alpha_g \to 0$ favors small values of most $\phi_{g,r}$ a-priori. Therefore, a data-dependent learning of α_g is essential in order to tune to the desired sparsity in $\boldsymbol{B}_g$. The subsection on hyperparameter specification discusses a model-based choice of α_g, along with the specific choices for $a_\lambda, b_\lambda, a_\tau$, and b_τ.

3.2.3 Bayesian Graphical Lasso Prior for modeling connectivity

Following Wang et al. (2012), to capture connectivity between different regions for individuals, $d_{i,g}$s are jointly modeled with a Gaussian graphical lasso prior. To be more precise,

$$\begin{aligned} &\mathbf{d}_i = (d_{i,1}, .., d_{i,G})' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^{-1}), \;\; i = 1, ..., n, \\ &p(\boldsymbol{\sigma}|\zeta) = C^{-1} \prod_{k<l} [DE(\sigma_{kl}|\zeta)] \prod_{k=1}^{G} \left[\mathrm{Exp}(\sigma_{kk}|\frac{\zeta}{2}) \right] \mathbf{1}_{\boldsymbol{\Sigma} \in \mathcal{P}^+}, \end{aligned} \tag{3.4}$$

where $\mathcal{P}^+$ is the class of all symmetric positive definite matrices and C is a normalization constant. $\boldsymbol{\sigma} = (\sigma_{kl} : k \leq l)$ is a vector of upper triangular and diagonal entries of the precision matrix $\boldsymbol{\Sigma}$. Using properties of the multivariate Gaussian distribution, a small value of σ_{kl} stands for weak connectivity between ROIs k and l, given the other ROIs. In fact, $\sigma_{kl} = 0$ $(k < l)$ implies that there is no connectivity between ROIs k and l, given the other ROIs. Thus, to favor shrinkage among off-diagonal entries of $\boldsymbol{\Sigma}$ for drawing inference on connectivity between pairs of ROIs, the Bayesian graphical lasso prior employs double exponential prior distributions on the off-diagonal entries of this precision matrix. The diagonals of $\boldsymbol{\Sigma}$ are assigned exponential distributions. Let $\boldsymbol{\eta} = (\eta_{kl} : k < l)$ be a set of latent scale parameters. Using the popular scale mixture representation of double

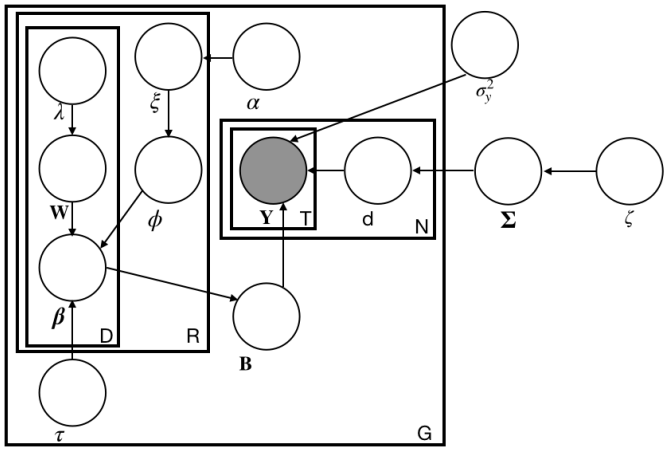

Figure 3.1: Plate Diagram Representation of the Proposed Model

exponential distributions (Wang et al., 2012), we can write

$$p(\boldsymbol{\sigma}|\zeta) = \int p(\boldsymbol{\sigma}|\zeta, \boldsymbol{\eta}) p(\boldsymbol{\eta}|\zeta) d\boldsymbol{\eta},$$

with $p(\boldsymbol{\sigma}|\zeta, \boldsymbol{\eta})$ given by

$$p(\boldsymbol{\sigma}|\zeta, \boldsymbol{\eta}) = C_{\boldsymbol{\eta}}^{-1} \prod_{k<l} \left[\frac{1}{\sqrt{2\pi\eta_{kl}}} \exp\left(-\frac{\sigma_{kl}^2}{2\eta_{kl}} \right) \right] \prod_{k=1}^{G} \left[\frac{\zeta}{2} \text{Exp}(-\frac{\zeta}{2}\sigma_{kk}) \right] \mathbf{1}_{\boldsymbol{\Sigma} \in \mathcal{P}^+}, \quad (3.5)$$

where $C_{\boldsymbol{\eta}}$ is the normalizing constant, which is an analytically intractable function of $\boldsymbol{\eta}$. The mixing density of $\boldsymbol{\eta}$ in the representation above is given by

$$p(\boldsymbol{\eta}|\zeta) \propto C_{\boldsymbol{\eta}} \prod_{k<l} \frac{\zeta^2}{2} \exp(-\frac{\zeta^2}{2}\eta_{kl}). \quad (3.6)$$

The hierarchy is completed by adding a Gamma prior on ζ, $\zeta \sim Gamma(a_\zeta, b_\zeta)$.

Finally, an inverse gamma prior $\sigma_y^2 \sim$ Inverse Gamma(a_σ, b_σ) is used on the variance parameter σ_y^2. A plate diagram of the model structure can be seen in Figure 3.1.

3.2.4 Hyperparameter Specification

The hyperparameters α_g in the stick-breaking construction are assigned a discrete uniform prior over 10 equally-spaced values in the interval $[R^{-D}, R^{-.10}]$, which will allow the data to dictate the level of sparsity appropriate for the prior (Guhaniyogi et al., 2017). The posterior updating of α_g under the griddy-Gibbs sampling algorithm can be found in section B.1. Although our extensive exploration with simulation data shows that a careful choice of fixed α_g value is able to produce inference as good as is offered by allowing α_g to vary, in practice, it is not clear how to choose such values. Consequently, a naive choice of α_g may offer incorrect conclusion from the model. On the other hand, the posterior updating of α_g in each iteration adds burden to computation. Therefore, this article allows posterior updating of α_g in the MCMC iteration, and fixes α_g values after the burn-in, at the value drawn in the n_{burn}-th iteration, where n_{burn} is the burn-in for the MCMC chain. As we discuss later, $n_{burn} = 100$ for simulation studies and equals 200 for the real data analysis. This both allows α_g to be learned while reducing any research bias stemming from a user dependent fixed value for α_g. The strategy works for various simulation studies and moderate perturbation of the prior range seems to produce robust inference. The values chosen for a_λ and b_λ have a strong effect on the shrinkage properties of the generalized double-Pareto prior, and setting $a_\lambda = 3$ and $b_\lambda = \sqrt[2p]{a_\lambda}$ prevents the prior for $\lambda_{g,j,r}$ from allowing for insufficient variance for $B_g[i_1, ..., i_D]$ to detect nonzero coefficients. Similar to Guhaniyogi et al. (2017), the hyperparameters a_τ and b_τ are set to $D - 1$ and $R^{1/D-1}$, respectively, in order to prevent overshrinkage with higher tensor response dimensions. Following Wang et al. (2014), a_ζ and b_ζ are set to 1 and 0.01, respectively, in order to preserve relative noninformativity of the Gaussian graphical prior. Finally, for both simulation studies and the real data analysis, a_σ

and b_σ were set to be 1 and $-\log 0.95$, respectively. While these hyperparameters are specified to provide readers a specific set of choices and they produce desirable results, we establish in section 3.4 that the inference is fairly robust with moderate perturbation of these hyperparameters.

3.3 Posterior Computation

The model framework and prior structure allow sampling from the posterior distribution using the Markov Chain Monte Carlo (MCMC) algorithm outlined in section B.1. In order to speed convergence of the MCMC chain, values for $\boldsymbol{\beta}_{g,j,r}$ are initialized using the singular value decomposition of the mode-j matricization of each $\hat{\mathbf{B}}_{g,MLE}$, the maximum likelihood estimate of the tensor-valued coefficient for the activation, assuming no connectivity component. This particular initialization method limits the rank R of the model to an upper bound of $\min_{g,j} p_{g,j}$. The posterior distributions of unknown quantities of interest are approximated by their empirical distributions from post burn-in MCMC samples.

Of particular interest in neuroscience is the assessment of whether a brain voxel is active or not, which, in our modeling framework translates to verifying whether $B_g[i_1, ..., i_D]$ is nonzero for voxel $(i_1, ..., i_D)$. It is well-acknowledged that the problem of selecting important cell coefficients is a challenging task when $\boldsymbol{B}_g$ is assigned a continuous shrinkage prior, since none of the cell coefficients is exactly zero in any MCMC iteration. The sequential 2-means method recently developed by Li and Pati (2017) was chosen over using posterior credible intervals to decide whether an element in $\mathbf{B}_g$ is significantly different from zero because it does not rely on choosing the size of posterior credible intervals through an arbitrary level of credibility. Bayesian multiple testing corrections proposed by Wacholder et al. (2004) and Whittemore (2007) provide analogs to the frequentist p-value that

address the false discovery rate, but still rely on setting an arbitrary threshold, which may be seen as undesirable for its subjectivity. Sequential 2-means, as outlined in Algorithm 1 is thus used to classify whether a coefficient is zero or nonzero from their post burn-in MCMC iterates. This is done separately for each reconstructed regional coefficient tensor $\mathbf{B}_g$. Following the suggestion in Li and Pati (2017), the value of b in Algorithm 1 is set to be the median of the standard deviations of the elements within $\mathbf{B}_g$. Within the estimates obtained through the use of the sequential 2-means method, nonzero-valued voxels are considered to be active.

Result: Final estimate of $\boldsymbol{\theta}$ with small elements set to be equal to 0

for $s \leftarrow 1$ **to** S **do**

Cluster the absolute value of elements in $\boldsymbol{\theta}^{(s)}$ into two clusters, $\mathcal{A}$ and $\mathcal{B}$, where $\bar{\mathcal{A}} \leq \bar{\mathcal{B}}$, where $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}$ denote the mean of elements in the clusters $\mathcal{A}$ and $\mathcal{B}$ respectively;

Cluster the elements of $\mathcal{A}$ into two clusters, $\mathcal{A}$ and $\mathcal{A}'$ such that $\bar{\mathcal{A}} < \bar{\mathcal{A}}'$;

while $|\bar{\mathcal{A}} - \bar{\mathcal{A}}'| > b$ **do**

Cluster the elements of $\mathcal{A}$ into two clusters, $\mathcal{A}$ and $\mathcal{A}'$ such that $\bar{\mathcal{A}} < \bar{\mathcal{A}}'$;

end

The number of elements remaining in $\mathcal{A}$ is the estimated number of true zero-valued elements, $n_z^{(s)}$, in $\boldsymbol{\theta}^{(s)}$

end

Find $\hat{n}_z$ = median value of n_z ;

Find $\hat{\boldsymbol{\theta}}$ = median values of the elements in $\boldsymbol{\theta}^{(1:S)}$;

Set elements in $\hat{\boldsymbol{\theta}}$ with the $\hat{n}_z$ smallest absolute values to 0

Algorithm 1: Sequential 2-means for posterior draws $s = 1, \ldots, S$ for parameter $\boldsymbol{\theta}$

In order to obtain an interpretable measure of the connectivity between re-

gions, the partial correlation between regions is examined. Since the partial correlation accounts for the correlation between two regions after removing the influence of all other regions (Das et al., 2017), it is expected to be an appropriate measure of pairwise connections. First, the partial correlation matrix was calculated from each posterior draw of $\mathbf{\Sigma}$ using the `prec2part` function in the `DensParcorr` package in R (Wang et al., 2018). Next, the sequential 2-means method (Li and Pati, 2017) was used on the upper-triangular elements of the partial correlation matrix to classify them as zero or nonzero. Regions with nonzero partial correlations are said to be connected (Warnick et al., 2018).

3.4 Simulation Studies

To validate the proposed methods, we simulate synthetic data with similar structure to that found in data collected from human fMRI studies. The tensor responses are simulated considering a block experimental design from the likelihood in (3.2). In each simulation study, we construct $G = 10$ different coefficient tensors corresponding to disjoint spaces, hereafter referred to as regions. For ease of visualization, the coefficient tensors are created to be three-dimensional, but can be generalized to any arbitrary dimension D so that the method may be applied to other scenarios. Throughout the simulation study, a sample size of $n = 20$ subjects is used, with the number of time points per subject being fixed at $T = 100$.

The covariate, $x_{i,t} = x_t$, was set to be the same for all of the subjects, without any loss of generality. A block experimental design is employed to generate the covariate, which consists of several discrete epochs of activity-rest periods, with the “activity" representing a period of stimulus presentations, and the “rest" referring to a state of rest or baseline. These activity-rest periods are alternated

throughout the experiment to ensure that signal variation, scanner sensitivity and subject movement have the similar effect throughout the experiment. To simulate activity-rest periods, we use the stimulus indicator function z_t as:

$$z_t = \begin{cases} 1, & \text{for } kP < t < kP + P/2, \quad k = 0, 1, \ldots \\ 0, & \text{otherwise} \end{cases}$$

for all t, given a defined period P for the block design. In our simulations, P is set to be 30. Next, the `canonicalHRF` function in the `neuRosim` package in R (Welvaert et al., 2011) is used to convolve the stimulus indicator z_t with the double-gamma canonical hemodynamic response function (HRF), which corrects for the expected delay between a stimulus and the resultant physiological response in the brain (Friston et al., 1998). This HRF is set using the default function values in `neuRosim` to have a delay of response relative to onset equal to 6 time steps, a delay of undershoot relative to onset of 12, a dispersion of response equal to 0.9, a dispersion of undershoot equal to 0.9, and a scale of undershoot equal to 0.35. The resulting covariate x_t is plotted in figure B.1. The dimensions of response tensor margins $p_{1,g}$, $p_{2,g}$, and $p_{3,g}$ all set to 10 for each region g, resulting in 10 regions with 1,000 voxels in each.

In order to demonstrate the effectiveness of the shrinkage component of the model, the true tensor coefficient values were randomly assigned using the `specifyregion` function from `neuRosim`. This function allows for the definition of tensors such that nonzero elements are spatially-contiguous spheres. In this simulation, the coefficient tensors are designed such that all elements took the value of either zero or 0.05. In real fMRI data, activation is typically observed in a small number of voxels/regions. Therefore we set the sizes of the true activated cells in our simulated data to be no greater than 5% of the total tensor size. The

true values for a slice of one of the coefficient tensors can be seen in Figure 3.2.

The contrast-to-noise ratio, defined as $B_{g,v}/\sigma_y$ for $B_{g,v} \neq 0$, was set to be equal to 0.05, which is proposed as a realistic value for neuroimaging data by Welvaert and Rosseel (2013). The connectivity between tensor regions was simulated by setting two pairs of the ten regions to have a region-wide correlation of 0.9, while all other regions were assigned correlations of zero. A covariance matrix ($\mathbf{\Sigma}^{-1}$) was created from this correlation matrix, and the region effects for subject i were simulated from a multivariate normal distribution with mean zero and covariance $\mathbf{\Sigma}^{-1}$. The signal-to-noise ratio, defined as $\Sigma^{-1}_{g,g'}/\sigma^2_y$ for $\Sigma^{-1}_{g,g'} \neq 0$ was set to 1, a realistic value based on Welvaert and Rosseel (2013). This quantity can be thought of as the relative effect of the connectivity on the observed response tensors. Finally, the observation-level variance (σ^2_y) was set to be 1.

3.4.1 Competitors

We fitted our proposed Bayesian model to the simulated data using different choices of rank R. In most of the real life applications, small values of R are sufficient to attain the desired inference, so models up to rank 7 were fit to the simulated data.

The performance of the proposed model is compared to that obtained from the following models: a vectorized model with a Generalized Double Pareto (GDP) shrinkage prior on the activation coefficients and a Gaussian graphical prior on the connectivity parameters, referred to as the *vectorized-GDP* approach; another vectorized model with a spike and slab prior on the activation coefficients referred to as the *spike-and-slab* approach; a general linear model (GLM). Details about these models are provided below.

The vectorized GDP model vectorizes the tensor response and builds a voxel

specific linear mixed effect model by regressing the response on predictors, followed by jointly estimating the voxel specific regression coefficients using a shrinkage prior distribution. More precisely, if Y_{g,i,t,v_1,v_2,v_3} is the response at voxel (v_1, v_2, v_3) in region g at time t for individual i, this mixed effect model proposes

$$Y_{g,i,t,v_1,v_2,v_3} \overset{ind.}{\sim} N(b^*_{g,v_1,v_2,v_3} x_{i,t} + d^*_{g,i}, \sigma^{*2}), \quad \boldsymbol{\beta}^*_g \sim \mathrm{N}(\mathbf{0}, \tau^*_g \mathbf{W}^*_g), \tag{3.7}$$

where $\boldsymbol{\beta}^*_g = (b^*_{g,v_1,v_2,v_3} : v_1 = 1 : p_{g,1}, v_2 = 1 : p_{g,2}, v_3 = 1 : p_{g,3})' \in \mathbb{R}^{p_{g,1} \times p_{g,2} \times p_{g,3}}$ is the vector of fixed effects and $\mathbf{W}^*_g = (w^*_{g,v_1,v_2,v_3} : v_1 = 1 : p_{g,1}, v_2 = 1 : p_{g,2}, v_3 = 1 : p_{g,3})$. The random effects $d^*_{g,i}$'s are jointly assigned a Gaussian graphical prior similar to (3.4). The hierarchical specification is completed by assigning $\tau^*_g \sim \mathrm{Gamma}(a_\tau, b_\tau)$, $w^*_{g,v_1,v_2,v_3} \sim \mathrm{Exp}\left(\frac{\lambda^{*2}_g}{2}\right)$, $\lambda^*_g \sim \mathrm{Gamma}(a_\lambda, b_\lambda)$. The vectorized GDP prior is a shrinkage prior on activation coefficients, which are envisioned as approximations to the Bayesian variable selection priors. Hence, as it is also the case with the spike-and-slab priors below, they take care of the multiplicity issues. Comparison with this vectorized-GDP reveals the advantage of retaining the tensor structure of the response to capture underlying local spatial structure while simultaneously inferring connectivity, as well as the advantage due to the parsimony offered by the PARAFAC decomposition. We also attempted to implement a spatially varying coefficient (SVC) model (Zhang et al., 2015) and found it to be extremely computationally demanding due to large matrix inversions in each MCMC iteration. Hence the comparison with SVC is not reported.

The spike-and-slab model utilizes the spike-and-slab prior, introduced in Equation (3.1) with $v_0 = 0.1$ for the "spike", or the part of the prior distribution with a high density around zero, and $v_1 = 10$ for the "slab", or the part of the prior distribution centered around zero, but with a much lower density around zero itself. The probability of inclusion in the spike prior component for voxel v, γ_v, was

assigned a Bernoulli prior distribution with probability π_v. Finally, each π_v was assigned a Beta$(2, 2)$ prior distribution. As shown by Scott and Berger (2010), keeping π_v random as opposed to fixed addresses the multiple correction issue in the spike and slab prior. As in the vectorized-GDP competitor, the random effects $d^*_{g,i}$ for the subjects and regions are assigned a Gaussian graphical prior as in (3.4).

Rather than using a cutoff to determine whether there is activation for a particular voxel within the coefficient tensor and for consistency in the comparison with the proposed approach, the sequential 2-means method outlined in Algorithm 1 is used on the posterior values of the coefficient tensor for the vectorized-GDP and spike and slab models. Similarly, the sequential-2 means method is also used to infer connectivity in both these competitors.

Finally, in order to improve the interpretability of the proposed model for those familiar with neuroimaging, the general linear model (GLM) is also used for comparison. It is of importance to note that the GLM is only placed on the activation components, as it fits each voxel to the covariate separately, which assumes that each voxel is independent of the others. In addition, sparsity in the activation cannot be built into the model assumptions, which has an effect on the uncertainty quantification of the tensor coefficient elements. The connectivity estimates do not have any uncertainty quantification at all. In order to determine activation, the Benjamini-Hochberg method for multiple testing correction (Benjamini and Hochberg, 1995) is used, as it allows for the control of the false discovery rate, which is desirable in high-dimensional regression settings. The residuals from the GLM are then summed within each region of interest and used to estimate the covariance between the regions. The covariance estimate is then used to estimate the precision and the partial correlation.

3.4.2 Comparison Metrics

MCMC is run for 1,100 iterations for all competitors, with a 100-iteration burn-in and the remaining used for inference. The assessment of convergence is made by the Raftery-Lewis diagnostic test implemented in the `R` package "coda". It shows a median effective sample size of 1,000 for the elements of all the $\mathbf{B}_g$s in the rank 1 model, and around 840 for the rank 2 through 7 models. Comparisons among competitors are based on a model fitting statistic and point estimation of $\boldsymbol{B}_g$'s. The accuracy of detecting active and inactive voxels for each competitor are also reported. Finally, we also compare competitors in terms of identifying connectivity between regions as the main goal of our proposed approach is to jointly detect activation and connectivity.

Model fitting is compared using the deviance information criterion (DIC), defined in Gelman et al. (2014) as $\text{DIC} = -2\log p(\mathbf{Y}|\hat{\mathbf{B}}, \hat{\mathbf{d}}, \mathbf{X}, \hat{\sigma}_y^2) + 2p_{DIC}$, where $p_{DIC} = 2\left(\log p(\mathbf{Y}|\hat{\mathbf{B}}, \hat{\mathbf{d}}, \mathbf{X}, \hat{\sigma}_y^2) - \frac{1}{S}\sum_{s=1}^{S}\log p(\mathbf{Y}|\mathbf{B}^s, \mathbf{d}^s, \mathbf{X}, \sigma_y^{2(s)})\right)$, $\hat{\theta}$ is the posterior mean of any parameter θ, S is the total number of post burn-in posterior samples. The superscript s denotes sth post burn-in posterior sample for a parameter, $\boldsymbol{Y}$, $\boldsymbol{X}$ are the collection of all responses and predictors respectively.

For comparison between the models in terms of point estimation of $\boldsymbol{B}_g$'s, we compute the square root of the mean squared error (RMSE) between the estimated tensor coefficient and true tensor coefficient, $\sqrt{\sum_{g=1}^{G}\sum_{v\in\mathcal{R}_g}(\bar{B}_{g,v} - B^0_{g,v})^2}$, where $\mathcal{R}_g$ represents region g, $B^0_{g,v}$ and $\bar{B}_{g,v}$ are the true and the posterior mean of the $v$the cell coefficient in the gth region respectively. In addition, as mentioned above, given the posterior mean estimates of $B_{g,v}$, sequential two-means approach (Li and Pati, 2017) is employed to identify active and inactive voxels. The true positive rate (TPR) and false positive rate (FPR) are computed for the different approaches. Finally, a summary of the performance of the connectivity for each

model is given as the Frobenius norm of the difference between the point estimate of the partial correlation ($\hat{\mathbf{\Omega}}$) and the true partial correlation matrix ($\mathbf{\Omega}$), with $\mathbf{\Omega} = \mathbf{\Sigma^{-1}}$. The Frobenius norm is defined for matrix $\mathbf{A}$ as $||\mathbf{A}||_F = \sqrt{\text{trace}(\mathbf{A'A})}$.

3.4.3 Results

In order to clearly define applications under which the proposed tensor model is expected to perform well, the following research hypotheses will be tested. Under the structure of the proposed tensor model, it is expected that sparse, hypercubic activation regions will be recovered well by the model. Sparse connectivity is also expected to be recovered effectively if the true partial correlation between regions is far enough away from zero. The model will also be tested in cases in which the observation error in the simulated data is strongly autocorrelated. The tensor models demonstrate benefit in terms of activation estimation, as seen in Figure 3.2. The slice of activation data shown was selected by finding the slice of one of the tensor coefficients that had the most true activation to clearly show the differences in the patterns of activation detection. The proposed model excels in its ability to identify activation only in voxels in and around where there is true activation, adequately capturing the localized spatial structure underlying the activation mechanism. This is an advantage over all other competitors. In particular, note that the spike-and-slab approach leads to a relatively large number of isolated false positives, while the vectorized-GDP identifies no active sites. The activation was also found using 99% posterior credible intervals, in which voxels within $\mathbf{B}_g$ with intervals that included zero were considered to be inactive. The plot created using this method can be found in appendix figure B.3.

Performance measures for all the competitors are summarized in Table 3.1. Using the DIC as a model selection criterion, the rank 3 model is chosen among

the proposed models. Overall, tensor models with rank greater than 1 outperform the vectorized-GDP in terms of sensitivity, with a sensitivity that is on par with the spike-and-slab. Due to the shrinkage imposed by the Bayesian models, the GLM has a better sensitivity and specificity for activation, however, it does not offer reliable uncertainty quantification, spatially localized activation, or reliable estimates of connectivity. Note that the spike-and-slab and vectorized-GDP competitors use far more parameters in their hierarchical models than the proposed tensor regression models to estimate the tensor coefficient. Therefore, the spike-and-slab and vectorized-GDP competitors are not compared with respect to the deviance information criterion, as we have found this criterion to be unreliable to compare models that have very large differences in the number of parameters due to underestimation of the penalty term. When viewing Table 1, it is important to keep in mind that we are interested in joint detection of activation and connectivity. Similar to the tensor coefficient, the sequential two-means method (Li and Pati, 2017) is used on the off-diagonal elements of the partial correlation matrix calculated from the precision matrix $\mathbf{\Sigma}$ to recover the connectivity structure among regions in the simulated data. In terms of the connectivity we see that the best performance is obtained by the tensor model of rank 3 and the worst is that obtained by the GLM approach. All unconnected regions are classified as having a partial correlation of zero, and the connected regions have nonzero partial correlations. The estimates from the model with the optimal rank as determined by the DIC and the competitor models are shown in Figure 3.3, which shows that the effect sizes are smaller than the true generative values. However, in settings with low signal-to-noise ratio and sample size, the proposed tensor model with rank 3 generally gives non-zero values for region pairs with true nonzero values, while leading to significant shrinkage in the true zero values. The GLM on the

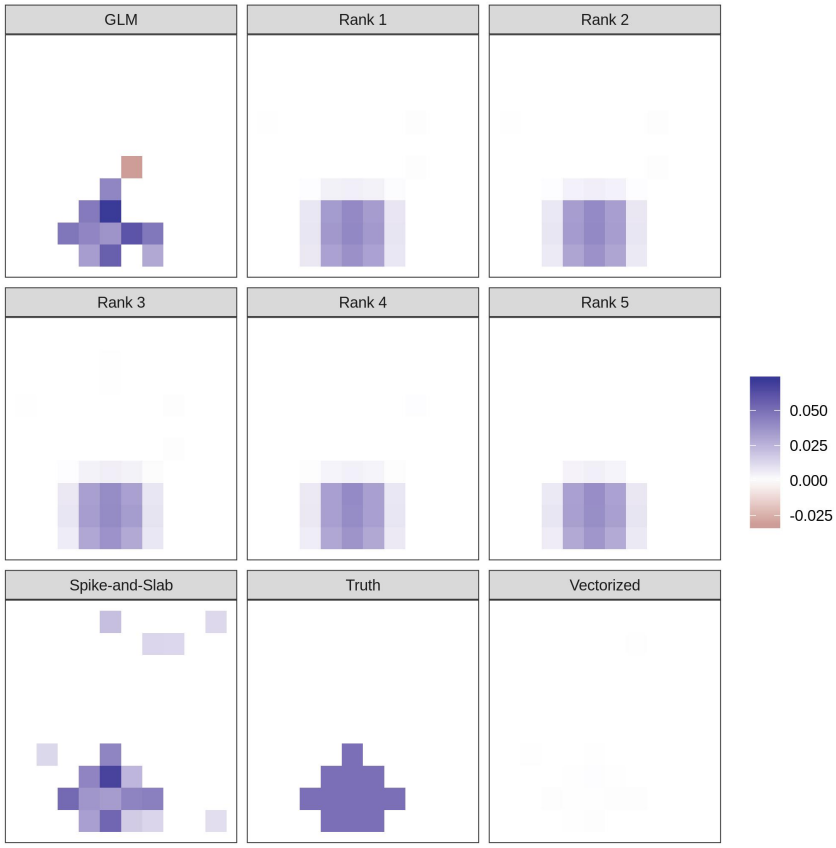

Figure 3.2: Rank model estimates and true value for a single slice of a three-dimensional coefficient tensor. Estimates are found using the sequential 2-means variable selection method (Li and Pati, 2017). The spike-and-slab and vectorized model estimates are also included for comparison.

other hand does not adequately estimate connectivity.

3.4.4 Sensitivity Analyses

Temporally correlated errors

Given that the model makes assumptions about independent errors, simulation tests were also conducted under data generation scenarios that violate this assumption. Therefore, in addition to having a scenario with independent errors $e_{g,i,t,\ell} \in \mathbf{E}_{g,i,t}$ as proposed in the original setting, we considered a new smaller scale

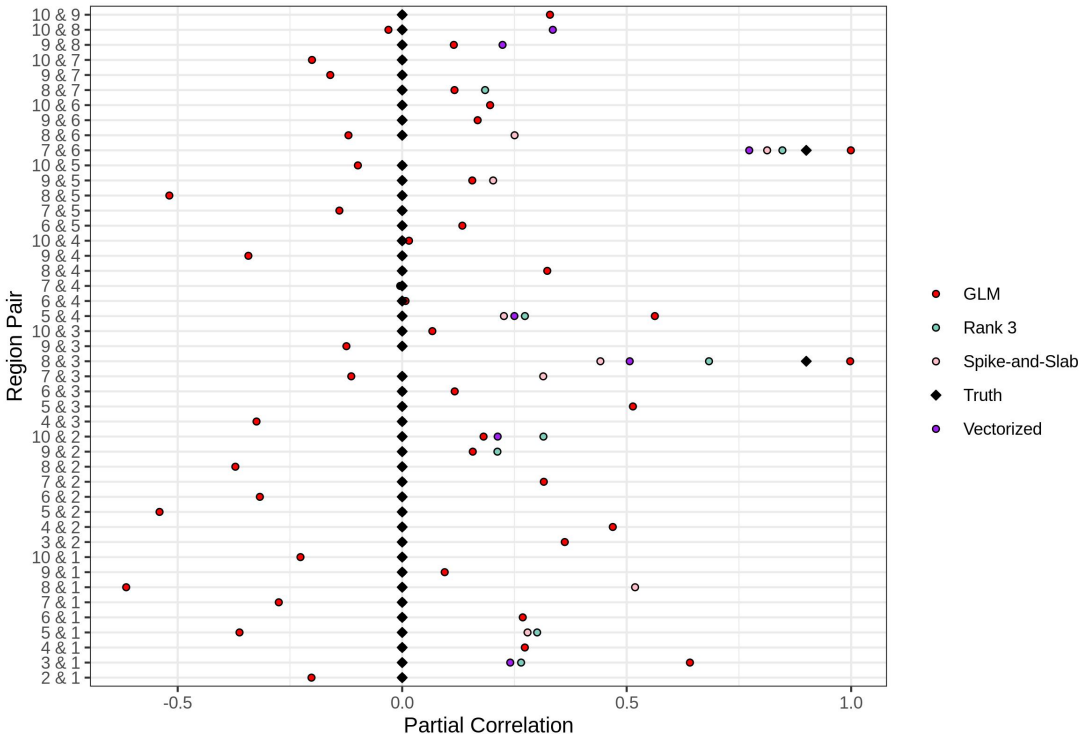

Figure 3.3: Estimates of the partial correlation for all possible region pairs after using the sequential 2-means method from Li and Pati (2017). The true partial correlation values for all region pairs are shown for comparison.

simulation scenario in which the errors have an autoregressive structure of order 1. That is, $e_{g,i,t,\ell} = 0.9e_{g,i,t-1,\ell} + u_{g,i,t,\ell}$, with $u_{g,i,t,\ell}$ assumed to be independent and identically distributed with mean zero and variance $\sigma_y^2 = 1$. This structure results in simulated data that are highly autocorrelated. In this new scenario we simulated $G = 5$ regions, each with response tensors of size $\mathbf{Y}_g \in \mathbb{R}^{20\times 20}$ for $n = 20$ subjects and $T = 100$ time steps. Furthermore, in order to directly show the effect of the autocorrelated errors on the model performance, an otherwise identical dataset of the same size was created in which the error in the data generation model was not autocorrelated.

For each of the two settings in this new simulation scenario, $\hat{\mathbf{B}}_g$ for the spike-and-slab competitor, the vectorized-GDP competitor, the GLM and the tensor

	# Parameters in $\mathbf{B}$	Time (Hrs)	DIC	RMSE for B
Rank 1	300	3.36	577863793	0.0247
Rank 2	600	4.08	577825593	0.0251
Rank 3	900	4.93	577724587	0.0252
Rank 4	1200	5.55	577877043	0.0252
Rank 5	1500	6.33	577823303	0.0252
Rank 6	1800	7.07	577849650	0.0252
Rank 7	2100	7.23	577806310	0.0252
Vectorized	10000	1.42		0.0232
Spike-and-Slab	10000	2.22		0.0241
GLM	10000	0.02		0.0037
	Sensitivity	Specificity	$\|\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|\|_F$	
Rank 1	0.4537	0.8124	1.1928	
Rank 2	0.6146	0.7896	1.0930	
Rank 3	0.6098	0.7891	0.9623	
Rank 4	0.6098	0.7911	0.9822	
Rank 5	0.6146	0.7934	1.1084	
Rank 6	0.6098	0.7945	1.1701	
Rank 7	0.6146	0.7954	1.1366	
Vectorized	0.5756	0.8871	0.9989	
Spike-and-Slab	0.6049	0.8696	1.2795	
GLM	0.8634	0.9955	2.7857	

Table 3.1: Performance diagnostics based on 1,100 draws from the posterior distribution with multiple different models using the same simulated data. For the performance measures of the Bayesian models, the first 100 draws from the posterior distribution are discarded as a burn-in.

model are created using the sequential 2-means method. The sensitivity and specificity were then found for each model, and can be seen in Table 3.2. It is important to note that the GLM does not do well in this scenario due to the low contrast-to-noise ratio combined with a smaller sample size in which only around 5% of the voxels in the coefficient tensor are actually nonzero. In spite of using the same multiple corrections method used for the GLM in the calculation of Table 3.1, in this case the GLM leads to a very poor sensitivity of 0.0283. Note that the specificity and sensitivity also go down for the other models also due to the smaller sample size, but the key part here is that the sensitivity measures are not affected by the induced temporal autocorrelation in the tensor models.

	Autocorrelated Error		Uncorrelated Error	
	Sensitivity	Specificity	Sensitivity	Specificity
Rank 1	0.5660	0.4836	0.5660	0.5290
Rank 2	0.5660	0.4182	0.5660	0.5090
Rank 3	0.5660	0.4134	0.5660	0.4979
Rank 4	0.5660	0.4113	0.5660	0.4836
Rank 5	0.5660	0.4097	0.5660	0.4667
Vectorized	0.5660	0.4266	0.5660	0.5026
Spike-and-Slab	0.5660	0.4261	0.5660	0.5048
GLM	0.0283	0.9974	0.0000	1.0000

Table 3.2: Comparison of Performance for Correlated and Uncorrelated Error

Contrast-to-noise comparisons

Next, we ran a set of scenarios in which the only change in the simulated data was the contrast-to-noise ratio, something that should have a significant impact on the activation inference. In each simulated dataset, again, the number of subjects was set to $n = 20$, the number of time steps was set to $T = 100$, the number of regions was set to $G = 5$, the signal to noise ratio was set to 5, and the observation noise was set to $\sigma_y^2 = 1$. Figure 3.4 shows the sensitivity and specificity for different values of the contrast-to-noise ratio. This shows that the proposed model is more sensitive than Bayesian competitors at low contrast-to-noise levels. This also shows that the higher-rank models have a higher specificity than the Bayesian competitors at higher contrast-to-noise ratios.

Hyperparameter sensitivity

Finally, in order to test the robustness of the model to choices of the hyperparameters, a grid of hyperparameter values was made by scaling each of the "standard" values for a_λ, b_λ, a_τ, b_τ, a_ζ, b_ζ, a_σ, and b_σ, defined in Section 3.2.4, by 0.01, 1, and 100, resulting in 6,561 different combinations. Of these, 100 settings were randomly sampled from the list and then tested with tensor model corresponding

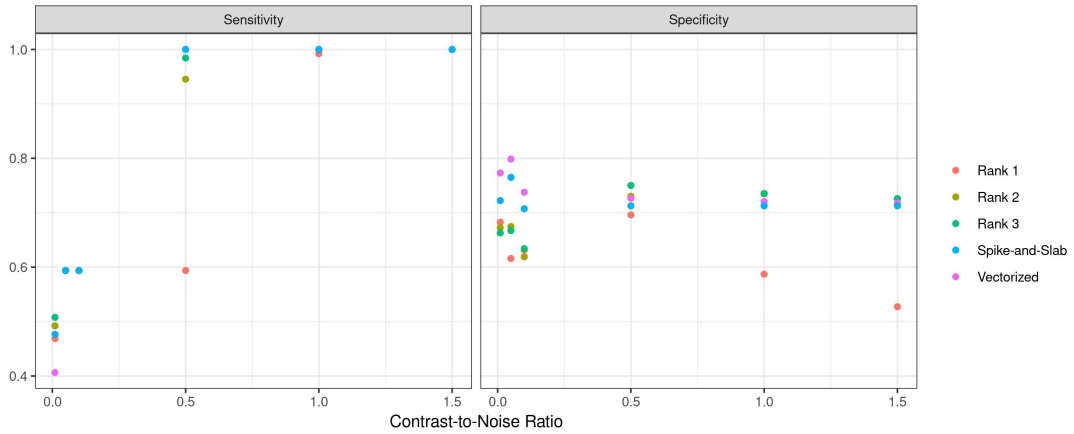

Figure 3.4: Sensitivity and specificity under varying contrast-to-noise ratios

to Rank 3. We graphed boxplots of the RMSE as well as length and coverage of 95% CI for all these hyperparameter combinations (these are available in figure B.2). The results are fairly robust with all three metrics varying within a small range under all different hyperparameter combinations. Overall, the simulation study reveals excellent recovery of activation and connectivity among regions by the proposed model. Although the computation time for the proposed model may a bit on the higher side, the burden is somewhat lessened by the rapid MCMC convergence for the model parameters allowing accurate inference even with a small burn-in.

3.5 Real Data Analysis

We return to the Balloon Analog Risk-Taking Task fMRI experiment that was described in section 2.5, now incorporating the data from multiple subjects in the experiment. Including multiple subjects allows for the inference from the model to be applied to more than one person, requiring different treatment of the data.

The preprocessing was done using FSL following what was done by Schonberg et al. (2012) as closely as possible. The fMRI have a repetition time (TR) of

2 seconds. In order to allow for T1 equilibrium effects, the first two scans were dropped. The EPI images were motion corrected, then high-pass filtered using a Gaussian least-squares linear fit with sigma = 50.0 seconds. Brain extraction was done using the BET function in FSL. The anatomic (T1-weighted) scans were registered using an affine transformation to standard Montreal Neurological Institute (MNI) space, and the EPI scans were then registered to each subject's corresponding anatomic scan. Finally, the data were spatially smoothed using a Gaussian kernel with a 5mm full-width half-maximum (FWHM). As these methods are implemented on the whole brain volumes all at once, the EPI scans were downsampled to have voxels with volume $8mm^3$ to find areas of increased activity within the entire brain in order to choose slices within the brain that can be analyzed at a higher resolution with voxels of volume $2mm^3$. For both cases, data were separated into 9 regions of interest based on the MNI structural atlas provided within FSL (Collins et al., 1995; Mazziotta et al., 2001). The MNI structural atlas is a hand-segmented atlas developed by Mazziotta et al. (2001) and Collins et al. (1995) and distributed within the FSL library of neuroimaging tools (Jenkinson et al., 2012). Choice of atlas is dependent on particular hypotheses, and can have a large effect on the connectivity inference of the proposed model. Splitting large regions of interest into two subregions has been found to be practically unnecessary, as the partial correlation within these subregions has been observed to be very high, even when the regions that were split apart are not physically contiguous within the brain. This does not present meaningful additional inference in the results of the model, so the regions of interest are kept as defined by the structural atlas in order to ease interpretation of the results. The voxel-level activation results were observed to be practically unchanged by splitting the regions of interest. One subject (subject 15) was removed from the dataset after

exploratory data analysis showed unusually high variance and temporal patterns that were not present in the scans of other subjects. In the whole-brain analysis, the regions of interest varied in size between 99 voxels and 1667 voxels after the BOLD response tensors were multiplied by binary masks, with a median region volume of 649 voxels.

To measure the level of risk being processed by a subject at a given time, we slightly modified the procedure used in Schonberg et al. (2012), described as follows. Begin with the centered number of pumps that an individual gave a âĂIJtreatmentâĂİ balloon before they âĂIJcashed-outâĂİ or the balloon exploded. It is assumed that the higher the number of pumps becomes, the more risk is present to the individual. This value was then convolved with the double-gamma haemodynamic response function (HRF), which takes into account the physiological lag between stimulus and response, and smooths the stepwise function for the centered number of pumps. In this analysis, all subjects are similar in age, and so they are assumed to have the same HRF. Future work may be done to expand the model to account for variance in the HRF for different subjects. The HRF used the default values in the `canonicalHRF` function in the `neuRosim` package in R, which are described in the Simulated Data Analysis section. Finally, we deviated from Schonberg et al. (2012) by subtracting the centered, convolved number of pumps on the control balloon from the treatment series to provide a basis of comparison between the two balloon types. To summarize, the final covariate is calculated as

$$\begin{aligned}\text{covariate} = &\\ &(\text{centered, convolved number of treatment pumps})\\ &-(\text{centered, convolved number of control pumps})\end{aligned}$$

Figure 2.8 shows the raw values for the centered number of control and treatment pumps, as well as the convolved pump functions and the final values for the covariate that were used in these analyses for a single subject.

The independent variable was then created as the difference between parametric modulation of the number of pumps on the treatment balloons and on the control balloons. The first two covariate values for each subject were removed to match the two dropped volumes in the EPI images. This covariate can be viewed as a contrast that accounts for the effect of the treatment balloons that mitigates any activation by subtracting the effect of increased pumps on the control balloon, when subjects are aware that there is no risk. This is similar to, but not the same as the analyses done in Schonberg et al. (2012) , which attempts to measure activity by subtracting effect estimates associated with the control balloons from the effect estimates associated with the treatment balloons. In this scenario positive coefficients imply more activity associated with treament balloons, negative coefficients suggest higher activity levels for the control balloons, and values close to zero imply no activity or similar activity for the treatment and control balloons.

The previous work done by Schonberg et al. (2012) concludes that areas within the frontal lobe, insula, and occipital lobe show BOLD response associations with risk-associated tasks. In this analysis, the data will be analyzed in order to verify these conclusions and explore the functional connectivity between the defined regions of interest. We hypothesize that our model will recover activations in the frontal lobe and insula, and that there will be some positive partial correlations between the two groups.

The proposed Bayesian tensor mixed effect models were fitted first on whole-brain data with low ranks in order to identify regions of the brain that should be examined further in a full-resolution analysis of a slice within the scans. This

is done in order to perform an analysis over the whole brain, which produces a dataset that does not fit into computer memory all-at-once at full resolution, without sacrificing precision on the estimate of voxel-level activation. Next, low-rank models were fitted to data from an axial slice in which $z = 18$ in the MNI standard space. This slice was identified as being within a region showing activity in the whole-brain analysis.

Due to the large size of the data, 2,200 samples were drawn from the joint posterior distribution of all of the parameters, and the first 200 samples were discarded as a "burn-in" measure. We believe that this is a sufficient posterior sample size based on examinations of the log-likelihood and autocorrelation functions, which can be seen in figures B.4 and B.5. In addition, the `effectiveSize` function within the `coda` package in R is used to calculate median values for the effective sample size for the 2,000 posterior draws of the elements in all $\mathbf{B}_g$ for the different rank models, see Table 3.3. This table indicates uncorrelated post burn-in posterior samples to draw reliable posterior inference.

The final estimates of the activation tensors were found following the sequential 2-means variable selection method as described in Li and Pati (2017), using the median posterior standard deviations of the elements within each tensor $\mathbf{B}_g$ as the tuning parameter. These estimates within the whole brain were then reorganized to their original positions, and can be seen for a single axial slice in Figure 3.5. Results for the high-resolution analysis of a single axial slice based on the whole-brain analysis can be seen in Figure 3.6. Higher values of the coefficient suggest that there is an increase in the BOLD response associated with higher levels of perceived risk. Larger positive values suggest that blood flow increases in these regions as risk increases. Larger negative values would suggest regions that exhibit a decrease in blood flow as risk increases, though no such regions were observed

in this analysis. The figures do show activations in the left posterior region of the frontal lobe and the anterior portion of the left insula, as concluded in Schonberg et al. (2012).

Similar to the simulation studies, the vectorized GDP, spike-and-slab, and GLM competitors are also fitted to the data to assess the advantages of preserving the tensor structure of the brain image in our proposed model. According to the Deviance Information Criterion (DIC) (Gelman et al., 2014) given in Table 3.3, Rank 2 is the best performing tensor mixed effect model for the higher resolution 2D slice data, while Rank 1 is the best performing model in the whole volume data. Figures 3.5 and 3.6 show that all the models that use tensor decompositions generally agree in terms of the posterior activation results. The vectorized model provides much lower estimates of activation strength than those obtained from the tensor decomposition models. These results suggest that the tensor mixed-effects models using the tensor decomposition is more sensitive than the GLM or vectorized GDP models, while also being more specific in detecting non-activation in regions further from active regions than the spike-and-slab model.

	Slice		Whole Brain	
	ESS	DIC	ESS	DIC
Rank 1	1826	1107186130	1743	1777033580
Rank 2	1742	1107078983	1583	1777227494
Rank 3	1742	1107160202	1211	1777227494
Rank 4	1709	1107130129		
Rank 5	1705	1107130129		
Spike-and-Slab	2000		2000	
Vectorized	1774		2000	

Table 3.3: The median effective sample size and log deviance information criterion for the five tensor decomposition models and a vectorized model for comparison.

The estimates for the significantly nonzero partial correlations between regions for the whole brain shown in Figure 3.7 indicate a functional connectivity

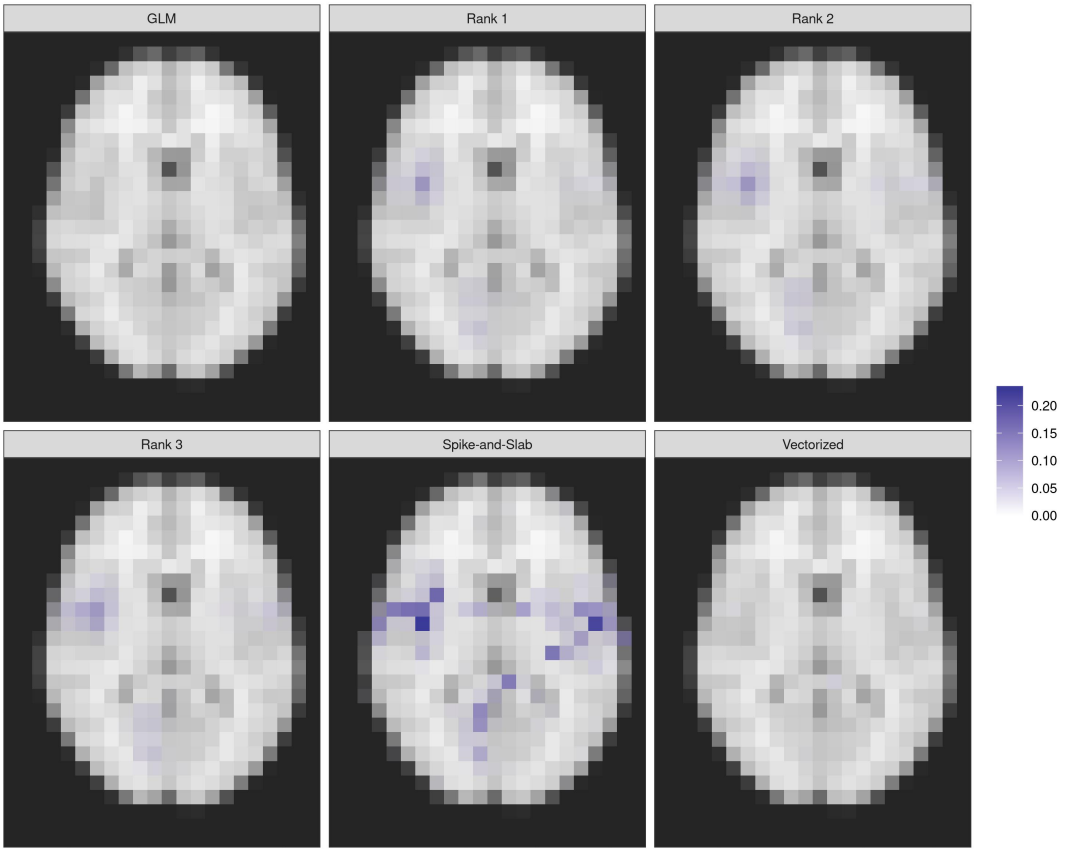

Figure 3.5: One Slice of Activity Estimates - Whole Brain

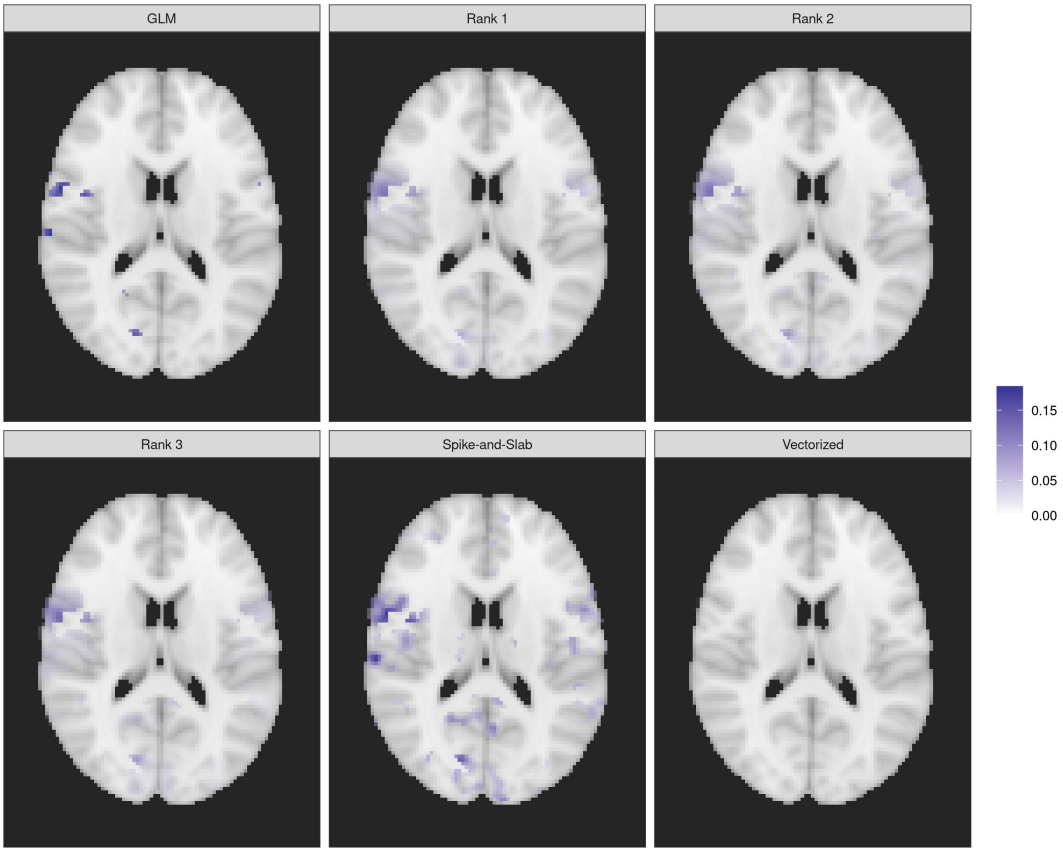

Figure 3.6: One Slice of Activity Estimates - Single Slice

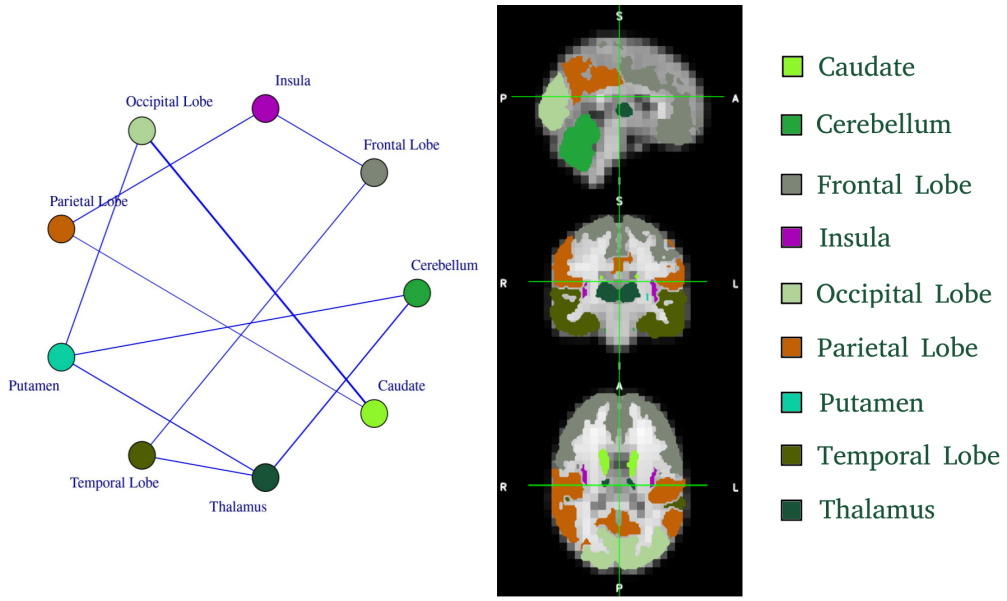

Figure 3.7: The connected regions of the whole brain in the rank 1 model, based on the partial correlation. The partial correlation here was found after using the sequential 2-means method (Li and Pati, 2017) on the partial correlation matrix elements across all MCMC samples. Thicker lines correspond to larger partial correlations, as all of the estmates of the nonzero partial correlations are positive.

network that also supports our hypothesis that the frontal lobe and the insula are connected. Other rank models lead to results with similar connective networks. The whole-brain connectivity network is shown because the results are more interpretable. Our finding agrees with earlier experiments suggesting that the frontal lobe plays a role in the assessment of risk (Miller and Milner, 1985). The numeric estimates for these partial correlations are small, which is to be expected in high-noise smoothed data, especially given the strength of the regularization in the Gaussian graphical prior (note that this prior induces strong shrinkage, see, for example the strength of the regularization in Figure 3.3 that is obtained in a much simpler simulation scenario with reasonable signal-to-noise data). However, these detected connectivity network in the regions of interest are significant and may be of investigative interest to neuroscientists.

3.6 Conclusions and Extensions

We present a new Bayesian tensor model for joint detection of voxel-level activation and region-specific connectivity in multi-subject studies. The proposed model produces markedly improved inference over a vectorized GDP model both in terms of identifying point estimation and quantifying uncertainty in a statistically principled manner. In addition, the model performs especially well in scenarios with low contrast-to-noise ratios, properly identifying hypercubic nonzero-valued regions within tensor coefficients while also finding functional connectivity between predefined regions. We also found that the proposed model exhibits an advantage over other Bayesian models in that nonzero estimates of activation tend to remain tightly clustered around true activation regions, preserving the underlying localized spatial structure. This is in contrast to both the spike-and-slab and vectorized GDP models considered as competitors. Our sensitivity analysis exhibited the robustness of the model to choices of hyperparameters. The proposed modeling structure is also flexible as it does not require extensive parameter tuning, adjustments for multiple testing, or selecting specific basis representations, making it accessible to a wide range neuroscientists and statisticians alike. Analysis of the model's performance under misspecification showed that it does not appear to be strongly impacted by the presence of temporal correlated errors. Due to the shrinkage priors imposed on the activation and connectivity components of the model, effect sizes are mildly underestimated, but activation and connectivity are still detected in low contrast-to-noise and signal-to-noise settings. Furthermore, our simulation studies show that the proposed tensor models performs better than competing models, particularly in comparison to the GLM, in terms inferring the connectivity structure across multiple regions.

Our analysis of a subset of the brain data examined by Schonberg et al. (2012)

confirmed that increased risk was associated with activity in the insula and frontal lobe. The inference was further improved by examining functional connectivity between regions of interest to detect a functional connectivity network, which we hope can be further explored in future research.

Our proposed approach assumes several important extensions. Notably, the parsimony in activation coefficients achieved by a PARAFAC decomposition may appear to be restrictive in certain applications, and can be replaced by a more flexible Tucker decomposition. Additional shrinkage prior models may also be explored under different tensor decomposition prior structures to explore the effects on posterior inference. Extensions of (3.2) that incorporate nonlinear regional effects through time could also be explored. Very importantly, extending the model to include temporal and spatio-temporal dependent error structures may further improve the model in its use with fMRI data. Finally, investigation into model-driven choices for subject-specific haemodynamic response functions may improve upon the accuracy of the proposed approach in real data applications.

Chapter 4

Bayesian Tensor Regression Using the Tucker Tensor Decomposition

Now we leave behind the modeling settings with one or more tensors as the response and examine a scenario in which each subject has a scalar response and tensor- and vector-valued covariates. This has implications about image analysis in general, but has features that are attractive to neuroimaging analyses in particular.

4.1 Introduction

Image analysis has become an important application area with the development of computer memory that allows for storing large datasets on local computing machines. Indeed, machine learning applications of image analysis are now present in many fields of research, such as medical imaging, character translation, and self-driving cars. In many of these cases, a fast modeling structure that provides inferential and/or prediction capabilities is all that is required. Such models have been shown to be very effective in settings with billions or trillions

of observed data points. However, scenarios in which models are built on smaller sample sizes suffer from a sensitivity to unusual observations without imposing additional constraints or assumptions.

The field of medical imaging is particularly rich in methods dealing with large datasets and small sample sizes. Since its inception with the discovery of X-rays in 1895 (Bercovich and Javitt, 2018), the field has grown to be a major component of modern medicine. The digitization of medical imaging in recent decades has opened the doors to analysts outside the purview of local hospitals and radiology centers, which lifts some of the burden of diagnosis from that of radiologists and reduces the rate of medical errors (Bruno et al., 2015). This shift also allows researchers to develop new models that can provide insight into how bodily mechanisms work inside living subjects by enabling the combined analysis over a number of subjects. Due to the clinical importance inference and prediction from these models, the assumptions and constraints must be carefully applied.

Several methods are already in use for these types of data within the neuroimaging community. One of the most commonly-used methods is referred to as the general linear model (GLM), which is not to be confused with the generalized linear model that is commonly used in statistics. This model performs a massive univariate analysis in which the response is regressed independently on each voxel within a tensor covariate, in addition to any additional vector covariates (Friston et al., 1995; Penny et al., 2011). These models have the advantage of being relatively inexpensive, computationally. However, they also assume that the associations between different voxels in the tensor and the response are all independent and not necessarily sparse. In practice, different multiple testing corrections are used to preserve spatial relationships among proximal voxels via independent components analysis, though work by Eklund et al. (2016) suggests

that these inflate the false discovery rate. Methods that control the false discovery rate are appealing (Benjamini and Hochberg, 1995; Lindquist and Mejia, 2015), but they fail to take spatial relationships within the tensor coefficient into account.

A different class of approaches takes advantage of the tensor structure of the data by decomposing the tensor covariates using one of two tensor decompositions and imposing regularization constraints. Work done by Zhou et al. (2013) uses the parallel factorization/canonical polyadic (PARAFAC/CP) tensor decomposition in a classical tensor regression approach, which assumes that dimension margins are principal components for construction of the tensor coefficients. This was expanded in the work by Li et al. (2018) in the use of the more flexible Tucker decomposition. Guhaniyogi et al. (2017) created a novel Bayesian prior structure on the PARAFAC/CP tensor decomposition elements, which improved on the uncertainty quantification from the model inference.

In this article, we will outline a model that satisfies the careful implementation of assumptions of sparsity and spatial similarity within a tensor-valued coefficient. This is accomplished through the use of the Tucker tensor decomposition (Tucker, 1966), which can be thought of as an analog to a principal components analysis where the components are dimension margins. Through simulation studies, we show the efficiency and accuracy of the model over a number of competitors, including other tensor regression models. The method is also applied in a neuroimaging analysis of data from the Alzheimer's Disease Neuroimaging Initiative.

4.2 Methodology

This section begins with the formulation of the Tucker tensor regression model and the priors applied to the components within it. The identifiability of the model

components is then discussed, followed by a discussion about the choice of rank in analyses. Competitor models are then outlined, which validate our proposed method as a useful and reliable tool in sparse tensor regression scenarios.

4.2.1 Tucker Tensor Regression Model

Assuming a scalar response y_i, for $i = 1, \ldots, n$ and vector-valued and tensor-valued predictors, $\boldsymbol{\eta}_i \in \mathbb{R}^q$ and $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, respectively, the observed linear model can be represented as

$$y_i = \langle \mathbf{B}, \mathbf{X}_i \rangle + \boldsymbol{\gamma}' \boldsymbol{\eta}_i + \epsilon_i, \tag{4.1}$$

in which $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is a tensor-valued coefficient, $\boldsymbol{\gamma} \in \mathbb{R}^q$ is a vector-valued coefficient, and ϵ_i is an error term, which follows any distribution centered at zero. Note that, since we are focusing on the methods to address tensor regression, we are restricting our attention to scenarios in which q is relatively small. However, the model framework here does allow for a seamless extension to a high-dimensional $\boldsymbol{\eta}_i$.

Given the potentially large predictor space within the tensor-valued covariate, it is reasonable to assume that some cases exist in which the association between the elements in $\mathbf{X}_i$ and y_i is sparse. In order to include this structure into the model while simultaneously reducing the parameter space, the Tucker tensor decomposition is used, as outlined in (1.2). Furthermore, in order to provide probabilistically-rigorous measures of uncertainty on the parameters in (4.1), Bayesian modeling is used. Specialized priors are imposed to induce sparsity on the estimates of the elements in $\mathbf{B}$. Standard Bayesian modeling techniques are applied to the $\boldsymbol{\gamma}$ and ϵ_i terms, as outlined below.

4.2.2 Prior structure

A key goal of the model is to address assumed sparsity in the tensor-valued coefficient. Classical regularization methods used to meet such a goal include various penalized regression algorithms like the LASSO (Tibshirani, 1996). However, these methods lack the ability to provide a measure of uncertainty quantification on the estimates for the parameters. Shifting to a Bayesian modeling structure can serve to fill this gap, leading to improved inference in some cases.

Li et al. (2018) proposes a classical model for tensor regression using the Tucker decomposition in which a penalty is applied to either the core tensor $\mathbf{G}$ or both the core tensor $\mathbf{G}$ and each dimension component $\boldsymbol{\beta}_{j,r_j}$ in the decomposition. We propose shrinkage priors on both the core tensor and all dimension components in order to more strongly induce parameter regularization.

Following the previous work done by Guhaniyogi et al. (2017), an adapted generalized double-Pareto prior is applied to the dimension components within the Tucker tensor decomposition. That is,

$$
\begin{aligned}
\mathbf{B} &= \sum_{r_1=1}^{R_1} \cdots \sum_{r_D=1}^{R_D} g_{r_1,\ldots,r_D} \boldsymbol{\beta}_{1,r_1} \circ \cdots \circ \boldsymbol{\beta}_{D,r_D}, \\
\boldsymbol{\beta}_{j,r_j} &\sim \text{Normal}\left(\mathbf{0}, \tau \mathbf{W}_{j,r_j}\right), \\
\tau &\sim \text{Gamma}(a_\tau, b_\tau), \\
\omega_{j,r_j,\ell} &\sim \text{Exponential}\left(\frac{\lambda_{j,r_j}^2}{2}\right), \\
\lambda_{j,r_j} &\sim \text{Gamma}(a_\lambda, b_\lambda),
\end{aligned}
$$

where $\mathbf{W}_{j,r_j}$ is a diagonal matrix with elements $\omega_{j,r_j,\ell}$ for $\ell = 1, \ldots, p_j$. Integrating over the element-specific scale parameters reduces the prior on $\beta_{j,r_j,\ell}$ to a double-

exponential distribution centered at 0 with a scale parameter of $\frac{\lambda_{j,r_j}}{\sqrt{\tau}}$, which has heavier tails than a Gaussian distribution, while also allocating higher densities around zero. This produces an analog to the adaptive LASSO, including its desirable oracle properties of the maximum a posteriori estimator (Armagan et al., 2013a).

In order to adequately select the proper rank for each dimension and to reduce noise in the tensor coefficient estimates, the generalized double-Pareto prior is also imposed on the elements of the core tensor $\mathbf{G}$:

$$\begin{aligned} g_{r_1,\ldots,r_D} &\sim \text{Normal}(0, z v_{r_1,\ldots,r_D}), \\ z &\sim \text{Gamma}(a_z, b_z), \\ v_{r_1,\ldots,r_D} &\sim \text{Exponential}\left(\frac{\varphi^2_{r_1,\ldots,r_D}}{2}\right), \\ \varphi_{r_1,\ldots,r_D} &\sim \text{Gamma}(a_\varphi, b_\varphi). \end{aligned}$$

This combination ensures that only summands within the Tucker decomposition that explain additional variance are included in the model. Selection of important ranks can be done using the sequential 2-means post-hoc variable selection procedure proposed by Li and Pati (2017).

For the purposes of the analyses in this article, a multivariate normal prior with mean $\boldsymbol{\mu}_\gamma$ and covariance $\boldsymbol{\Sigma}_\gamma$ is placed on the elements of $\boldsymbol{\gamma}$. This is done to maintain conjugacy while maintaining control over the expected effects of the vector-valued coefficients. The errors ϵ_is in this model are assumed to be independent identically distributed following a normal distribution with a mean of zero, and a variance of σ^2_y. An inverse gamma prior is placed on σ^2_y with hyperparame-

ters a_σ and b_σ.

$$\begin{aligned}
\boldsymbol{\gamma} &\sim \text{Normal}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma) \\
\epsilon_i &\overset{i.i.d.}{\sim} \text{Normal}(0, \sigma_y^2) \\
\sigma_y^2 &\sim \text{Inverse Gamma}(a_\sigma, b_\sigma)
\end{aligned}$$

4.2.3 Identifiability

The decomposition of the tensor coefficient raises questions about the identifiability of the parameters in the model. That is, if the value of any voxel within the tensor coefficient is estimated as

$$\hat{b}_v \in \hat{\mathbf{B}} : \hat{b}_v = g_{1,\ldots,1}\beta_{1,1,v_1} \cdots \beta_{D,1,v_D} + \ldots + g_{R_1,\ldots,R_D}\beta_{1,R_1,v_1} \cdots \beta_{D,R_D,v_D},$$

where v is the voxel location within the tensor ($v = (v_1, \ldots, v_D)$), then any two of these summands can be multiplied by c and $\frac{1}{c}$, respectively, and the estimate for that voxel within the tensor coefficient $\hat{b}_v$ would have the same value. Indeed, the values of $g_{r_1,\ldots,r_D}$ and β_{j,r_j,v_j} are not identifiable, but $\hat{b}_v$ remains identifiable through the shrinkage priors on the tensor decomposition components.

Something to consider when using this model for a data analysis is that an identifiability problem exists between $\mathbf{B}$ and $\boldsymbol{\gamma}$ if there are voxels within $\mathbf{X}$, $\forall i = 1, \ldots, n$ that are collinear with any of the values in $\boldsymbol{\eta}$, $\forall i = 1, \ldots, n$. This is a problem that may be undetected in two-step frequentist models with regularization, as either elements in $\mathbf{B}$ or $\boldsymbol{\gamma}$ may simply be assigned values of zero.

4.2.4 Selection of Rank

A key consideration in the use of the proposed model structure is the selection of each dimension rank. Increases in a margin's rank can be made with an attempt to increase the spatial resolution on the inference of the tensor coefficient. For larger tensor covariates, this may require higher ranks than smaller tensors if the nonzero coefficients are not hypercubic. Choice of unequal ranks may be prudent if some of the tensor dimensions are much larger or smaller than others. Consider an example in which the tensor covariate has dimensions $100 \times 100 \times 4$. The ranks for the first two dimensions may need to be considerably larger than the rank for the final dimension, as there are only four possible margin locations in the final dimension. One such realistic application would be to combine magnetic resonance images that use different sequences (e.g. T1 weighted, T2 weighted, effective T2, etc.) into a single tensor, using a low rank on the dimension that represents different sequences. This is a clear advantage over the CP/PARAFAC decomposition methods, as each dimension is not forced to have the same number of ranks as all of the others.

Once a set of reasonable ranks is decided for each tensor dimension, the different combinations of the ranks should be used to fit models and compare results using some model selection criterion like the DIC.

4.2.5 Competitor Models

The effectiveness of the proposed Bayesian sparse tensor regression models is shown by making direct comparisons to models commonly used in the field of neuroscience. The first is the general linear model (GLM), in which the response $(y_1, \ldots, y_n) = \mathbf{y}$ is regressed on each voxel v within the tensor covariate $\mathbf{X} \in \mathbb{R}^{p_1,\ldots,p_D}$ independently with the vector-valued coefficients. The linear model for

the GLM is written in equation (4.2). Note that the same responses $(y_1, \ldots, y_n) = \mathbf{y}$ are used to fit separate models for each voxel within the tensor covariate.

$$y_i = b_v X_{i,v} + \boldsymbol{\gamma}' \boldsymbol{\eta}_i + \epsilon_{i,v} \tag{4.2}$$

This model is an industry standard for its ease of implementation and rapid completion. However, this model often suffers from a high false discovery rate (Eklund et al., 2016), and does not provide a single estimate for the vector-valued coefficient. We adjust for the high false discovery rate in our implementation by using the Benjamini-Hochberg multiple testing correction (Benjamini and Hochberg, 1995), which fixes the false discovery rate (FDR). In the following trials, the FDR is fixed at 0.05, in agreement with standard practice in neuroimaging. In order to provide a single estimate for the vector of coefficients $\boldsymbol{\gamma}$, a two-step GLM algorithm could be used, which is described in algorithm 2. In the implementation of these models, δ is set to be equal to 1.

Result: Estimates of $\boldsymbol{\gamma}$ and $\mathbf{B}$

Initialize $\boldsymbol{\gamma}^{(0)}$ as $\max_{\gamma} \ell(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma})$, $\mathbf{B}^{(0)}$ as $\max_{\mathbf{B}} \ell(\mathbf{y}, \boldsymbol{\eta}, \mathbf{X}, \boldsymbol{\gamma}^{(0)}, \mathbf{B})$;

repeat

Set $\boldsymbol{\gamma}^{(t+1)} = \max_{\gamma} \ell(\mathbf{y}, \boldsymbol{\eta}, \mathbf{X}, \boldsymbol{\gamma}, \mathbf{B}^{(t)})$;

Set $\mathbf{B}^{(t+1)} = \max_{\mathbf{B}} \ell(\mathbf{y}, \boldsymbol{\eta}, \mathbf{X}, \boldsymbol{\gamma}^{(t+1)}, \mathbf{B})$;

until $\max\left(\max(\mathbf{B}^{(t+1)} - \mathbf{B}^{(t)}), \max(\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\gamma}^{(t)})\right) < \delta$;

Algorithm 2: Two-step method for regressing response $\mathbf{y}$ on vector-valued covariates $\boldsymbol{\eta}$ and tensor-valued covariates $\mathbf{X}$.

The advantage of the two-step GLM is that it produces single estimates of $\boldsymbol{\gamma}$, and as a result, tends to have a lower FDR. Here again, the Benjamini-Hochberg multiple testing correction is used to limit the false discovery rate in deciding which voxels in the tensor coefficient are significantly different from zero.

Direct tensor regression competitors are also used to validate the performance

of the proposed model structure. Specifically, the frequentist tensor regression using the PARAFAC/CP tensor decomposition (Zhou et al., 2013) and the Tucker tensor decomposition (Li et al., 2018) are used to compare point estimates to classical methods. Comparison to the Bayesian tensor regression using the PARAFAC/CP tensor decomposition (Guhaniyogi et al., 2017) provides a more direct comparison in terms of point estimates and uncertainty quantification.

4.3 Simulated Data Analysis

On order to demonstrate the efficacy of the proposed Bayesian model, data were simulated from the linear model in (4.1) under the conditions $\mathbf{B} \in \mathbb{R}^{50\times 50}$, where nonzero-valued elements take the value of 1 in the middle of the regions, fading outward to lowere nonzero values using the `specifyregion` function within the `neuRosim` package in R (Welvaert et al., 2011). For the sake of these simulations, three separate regions of activation are generated at random locations. The elements of $\mathbf{X}_i$ are all independently generated from a standard normal distribution for $i = 1, \ldots, 1000$. The elements for the vector-valued covariates $\boldsymbol{\eta_i}$ are also independently generated from a standard normal distribution. The parameters for the vector-valued covariates are set as $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3) = (25, 3, 0.1)$ to show how the different models estimate parameters of different size, relative to the observation error, which is set to have a variance of 1. Finally, the elements of $\mathbf{y} = (y_1, \ldots, y_n)$ are generated according to (4.1) for $\epsilon_i \sim \text{Normal}(0, 1)$.

The Bayesian models had the following hyperparameter settings: $a_\sigma = 3$, $b_\sigma = 20$, $a_\lambda = 3$, $b_\lambda = a_\lambda^{1/(2D)}$, $\boldsymbol{\mu}_\gamma = \mathbf{0}$, $\boldsymbol{\Sigma}_\gamma = 900\mathbf{I}_q$, $a_\tau = 1$, $b_\tau = \min(R_1, \ldots, R_D)^{(1/D)-1}$, $a_z = 1$, $b_z = \min(R_1, \ldots, R_D)^{(1/D)-1}$, $a_\varphi = 3$, and $b_\varphi = a_\varphi^{1/(2D)}$. The prior on the observation error variance is set to be relatively noninformative in the context of the simulation, with a mean of 10 and a variance of 100. The priors for λ and φ

are set to have modes between 1.5 and 2, getting closer to 2 as the tensor dimension increases, which places the prior expected value for $v_{r_1,\ldots,r_D}$ and the elements within $\mathbf{W}_{j,r_j}$ to be between $\frac{2}{3}$ and $\frac{1}{2}$. The priors for τ and z have a mean of 1 when all $R_1,\ldots,R_D = 1$, increasing sublinearly with both the minimum value of R_j for $j = 1,\ldots,D$, and the tensor dimension D. The prior variance for τ and z increases linearly in rank when $D = 2$, and superlinearly in rank when $D > 2$. This prior specification allows for a slightly higher prior variance for $\boldsymbol{\beta}_{j,r_j}$ and $g_{r_1,\ldots,r_D}$ as rank and dimension increase in order to allow for moderately faster exploration of the parameter space as the tensor dimension and rank increase. The prior for $\boldsymbol{\gamma}$ is set to be relatively noninformative.

For all 16 models with $R_1, R_2 \in \{1,2,3,4\}$, 11,000 Markov Chain Monte Carlo (MCMC) simulations from the posterior distribution were drawn. After discarding the first 1000 draws, the remaining draws were used to estimate $\mathbf{B}$ and $\boldsymbol{\gamma}$.

Since the model does not produce exact zeros as estimates for zero-valued parameters, the sequential 2-means variable selection method proposed by Li and Pati (2017) is used for variable selection within the tensor coefficient. The method works by using 2-means clustering of any subset of the parameter space on the absolute values of each draw from the posterior distribution. The number of values in the cluster in which the center is furthest from zero is taken as the number of non-zero valued parameter estimates for that particular posterior sample. The median number of non-zero parameter estimates across all of the posterior samples, m is then found. Finally, the parameters with the m highest posterior median absolute values are determined to have true non-zero values. These parameters are then estimated with their posterior medians. This process is formally described in algorithm 1.

Point estimates for the Bayesian and frequentist sparse Tucker tensor regres-

sion and the Bayesian and frequentist sparse CP tensor regression can be seen in Figure 4.1. In the interest of clarity, the only Tucker decomposition models shown are the equivalents to the PARAFAC/CP models. It is worth noting that additional ranks are needed in order to accurately detect additional nonzero regions in the true coefficient tensor. The Bayesian methods using the CP/PARAFAC decomposition do not exhibit the same gains in nonzero region detection as the Tucker decomposition models. All of the tensor regression methods show significant improvement over the two-step GLM competitor. In order to choose between the models, the deviance information criterion is used for the Bayesian sparse tensor regression models, while the log-likelihood is used for the frequentist sparse tensor regression models. The point estimates from these selected models, along with the two-step GLM estimate and the true values are shown in figure 4.2. The estimate shown for the two-step GLM tensor coefficient assigns values of zero to voxels within the tensor that are not statistically significant after using the Benjamini-Hochberg multiple testing correction.

In order to compare the performance of these models, the root mean squared error (RMSE) for each element within $\mathbf{B}$ can be compared in Table 4.2. Explicitly, the RMSE was found as $\sqrt{\frac{1}{V}\sum_v(\hat{B}_v - B_v)^2}$, where $\hat{B}_v$ is the point estimate for tensor coefficient element v obtained from a model, B_v is the true value taken by the coefficient tensor element v, and V is the total number of elements in the coefficient tensor. These results show the effectiveness of the Bayesian sparse Tucker tensor regression model in estimating true tensor coefficient values, exhibiting a lower RMSE than other tensor regression methods using tensor decompostions. In particular, significant improvements are seen over the Bayesian CP/PARAFAC models when both R_1 and R_2 are greater than 1. This is also visible in figure 4.1, where the gain in the true image recovery is much greater going from a rank 1,1

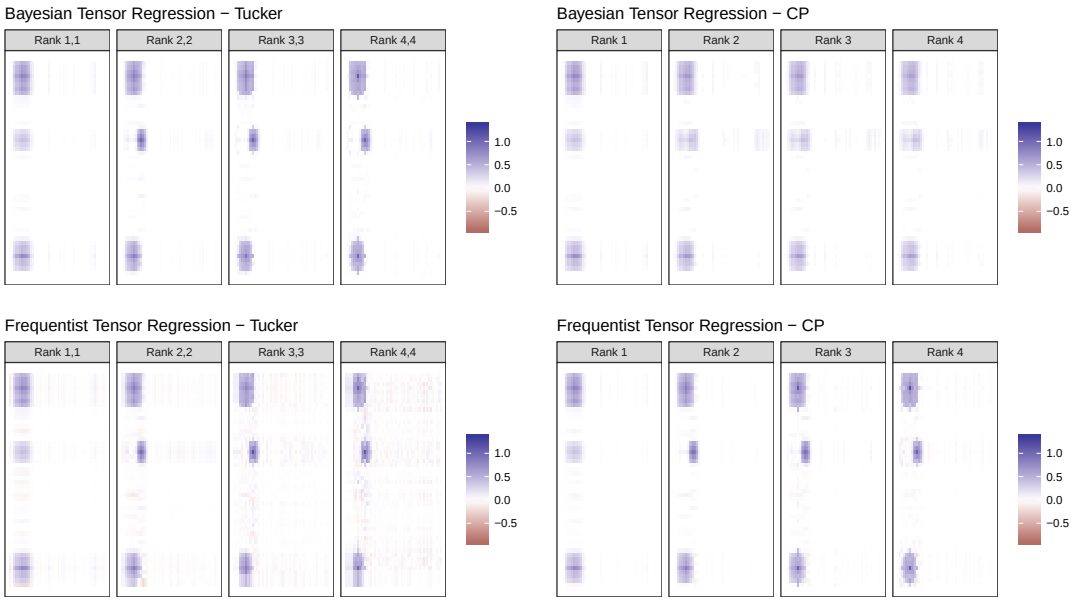

Figure 4.1: Point estimates of the true tensor coefficient for the Bayesian sparse Tucker tensor regression and the frequentist sparse Tucker tensor regression. The Bayesian point estimate was found using the sequential 2-means posterior variable selection method (Li and Pati, 2017).

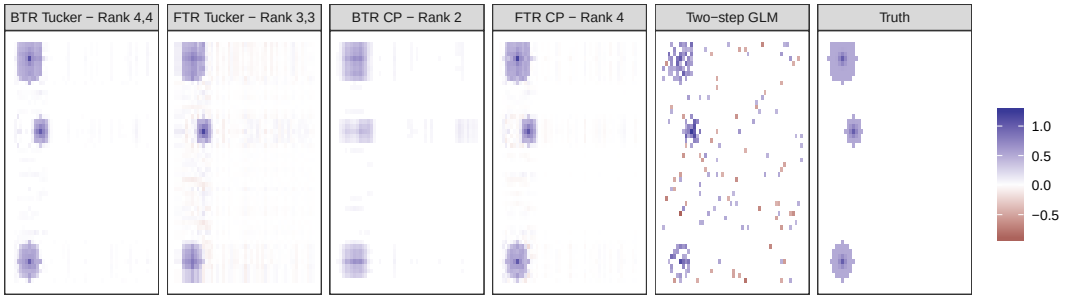

Figure 4.2: Point estimates of the true tensor coefficient for all competitors. Bayesian models were chosen using the deviance information criterion, and the frequentist models were chosen using the Bayesian Information Criterion.

	True Zero			
R_1/R_2	1	2	3	4
1	0.955	0.955	0.955	0.955
2	0.955	0.965	0.965	0.966
3	0.959	0.966	0.968	0.970
4	0.959	0.965	0.971	0.975
	Rank 1	Rank 2	Rank 3	Rank 4
BTR CP	0.968	0.991	0.994	0.998
	True Nonzero			
R_1/R_2	1	2	3	4
1	0.391	0.384	0.391	0.404
2	0.391	0.503	0.503	0.503
3	0.404	0.503	0.669	0.656
4	0.391	0.503	0.642	0.735
	Rank 1	Rank 2	Rank 3	Rank 4
BTR CP	0.464	0.603	0.636	0.695

Table 4.1: Coverage probabilities of the 95% credible intervals for the true zero and true nonzero values within **B** for different rank models.

to rank 2,2 model than it is in the Bayesian CP/PARAFAC models going from Rank 1 to Rank 2. In addition, all of the competitor methods perform significantly better than the two-step GLM model, which is often used in neuroimaging studies.

The posterior densities for $\{\gamma_1, \gamma_2, \gamma_3\}$ under different models can be seen in Figure 4.3. These results show that the posterior densities for each γ_j are centered at or around the true values. When a coefficient is very far from zero relative to the error in the likelihood distribution, as in the first column of plots in figure 4.3, the Bayesian models do show the effect of choosing a prior mean of 0. This inference would be improved by choosing prior mean values that more accurately reflect what is expected to be seen within the data, but the effect of this poorly specified prior are not extreme. However, when the true value is very close to zero, as in the last column of plots in figure 4.3, the Bayesian models still have

	BTR Tucker			
R_1/R_2	1	2	3	4
1	0.0686	0.0686	0.0686	0.0686
2	0.0686	0.0413	0.0413	0.0413
3	0.0686	0.0413	0.0330	0.0330
4	0.0687	0.0413	0.0330	0.0276
	FTR Tucker			
R_1/R_2	1	2	3	4
1	0.0718	0.0697	0.0719	0.0752
2	0.0711	0.0467	0.0488	0.0544
3	0.0719	0.0475	0.0501	0.0591
4	0.0725	0.0491	0.0506	0.0617
	CP/PARAFAC			
Model	Rank 1	Rank 2	Rank 3	Rank 4
BTR	0.0690	0.0614	0.0614	0.0633
FTR	0.0686	0.0420	0.0383	0.0297
	No Ranks			
GLM	0.1355			

Table 4.2: The root mean squared error (RMSE) for the estimates of $\mathbf{B}$ under Bayesian sparse tensor regression (BTR) and frequentist sparse tensor regression (FTR) using the Tucker and CP tensor decompositions. The RMSE for the general linear model (GLM) is provided as a comparison.

high posterior densities above 0, suggesting that they do detect the weak signal for a small true coefficient value. Interestingly, the Bayesian sparse Tucker tensor regressions show much higher posterior densities around the true coefficient values in all cases. This is an effect of the improved inference estimating $\mathbf{B}$.

4.3.1 Model Convergence

In all Bayesian modeling settings, it is important to ensure that the MCMC converges around an area within the neighborhood of the global posterior mode. Given that the posterior inference matches the inference shown within the fre-

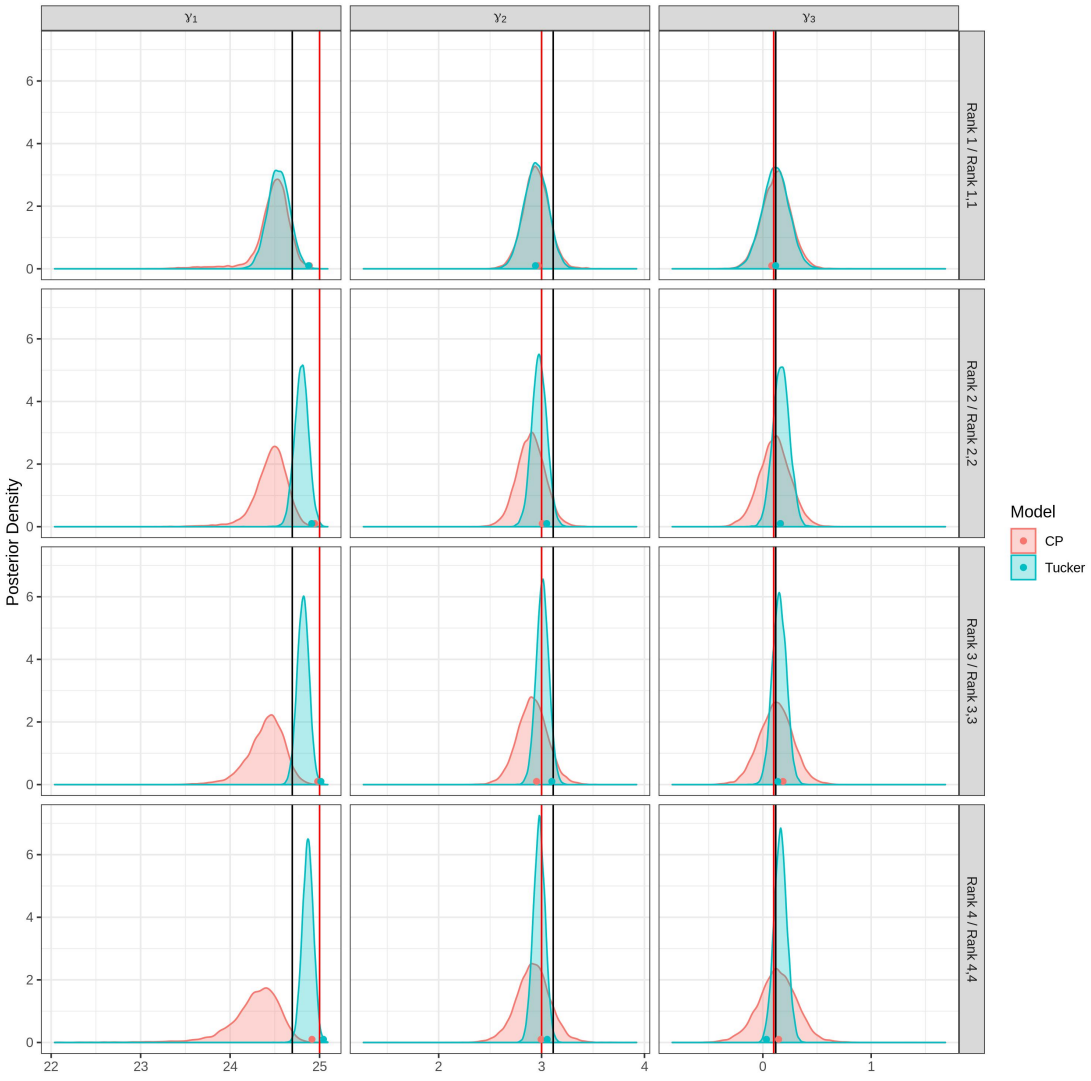

Figure 4.3: Posterior densities for $\{\gamma_1, \gamma_2, \gamma_3\}$ under different models. Points indicate the estimates from the frequentist sparse tensor regression models. The red line indicates the true value, and the black line indicates the estimate from the two-step GLM frequentist model.

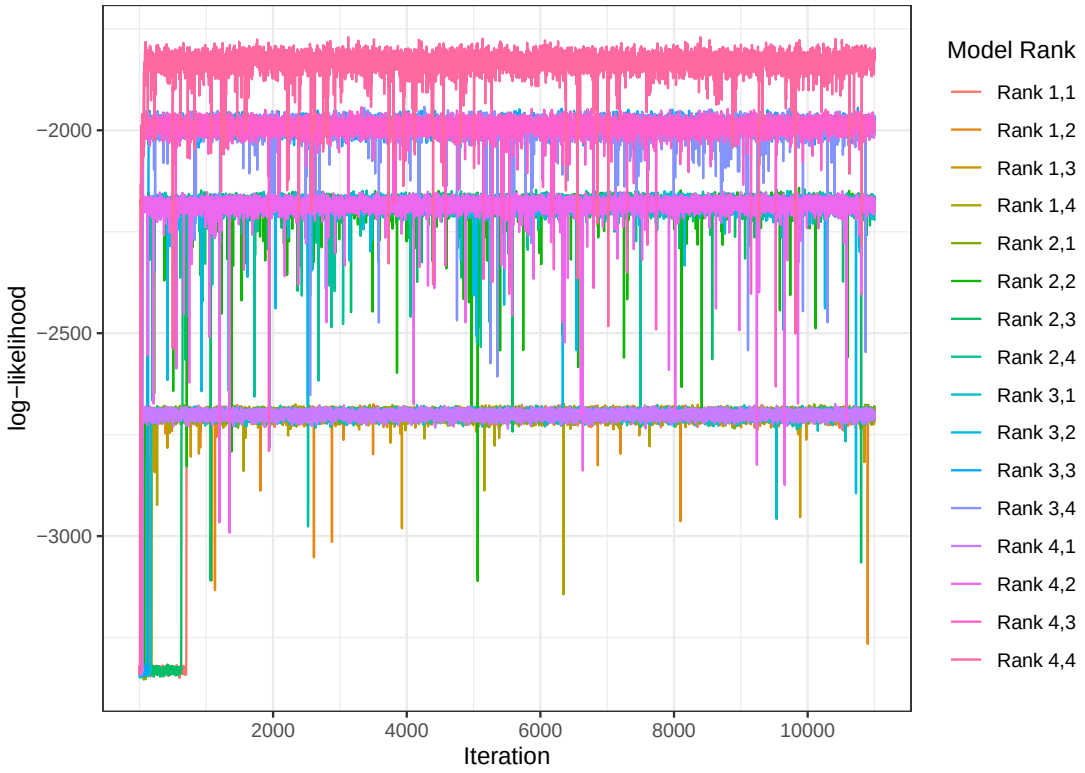

Figure 4.4: The log-likelihoods for the models applied to the simulated data

quentist models, the model does converge to a global mean in the simulated data settings when the assumption of normal error is satisfied, and the prior distributions are specified to be relatively uninformative. The log-likelihood values using the posterior draws, shown in Figure 4.4, also show rapid convergence to a mode and posterior stability.

4.3.2 Hyperparameter Sensitivity

Modeling in such a high-dimensional space with several hierarchical levels in the prior structure requires careful selection of hyperparameter values. However, the model is still expected to be somewhat robust to differences in prior specification. In order to test this expectation, a large grid of hyperparameter values

for $a_\sigma, b_\sigma, a_\lambda, a_\tau, a_z$, and a_φ was created. First, the values for the hyperparameters used to model the simulated data are taken and multiplied by 0.1, 1, or 10. One hundred configurations were randomly selected from the 729 possible configurations, and the model with ranks $R_1 = R_2 = 3$ was run with the same simulated data as above for 11,000 iterations with each combination. In each case, 1,000 iterations in the MCMC were discarded as a burn-in.

Of the 100 configurations, 93 converged to have an RMSE for the tensor coefficient $\mathbf{B}$ between 0.03279 and 0.03336. The remaining 7 configurations all had RMSEs within 10^{-8} of 0.1336523, which is the RMSE that results from predicting $\mathbf{B}$ with a tensor of all zeros. The priors for these seven configurations all had the smallest values for the prior variance for the elements of $\mathbf{B}$ (approximately 2×10^{-3}) and/or the highest prior means (approximately 10^{10}) and variances (over 10^{20}) for σ_y^2. Indeed, individual inspection of these models with larger RMSE values show that the estimates for $\boldsymbol{\gamma}$ are within the same range as the estimates from the models that had a lower RMSE for the tensor coefficient, but they all estimate that the tensor coefficient is zero. This highlights the need to carefully balance the amount of shrinkage that can be applied to elements within $\mathbf{B}$ before no image coefficients are recovered. In addition, the prior specification for σ_y^2 must be reasonable in order to achieve interpretable results.

4.4 Neuroimaging Analysis

Data from magnetic resonance images (MRIs) of the brain have been found to have sparse, spatially-contiguous associations with certain phenotypes, such as cognitive performance scores or neurological disorders. In order to demonstrate the efficacy of the Bayesian sparse Tucker tensor regression, data from the AlzheimerâĂŹs Disease Neuroimaging Initiative (ADNI) database (`adni.loni.`

usc.edu) were used. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AlzheimerâĂŹs disease (AD). For these analyses, data from the "ADNI1: Complete 1 Yr 1.5T" image collection were used. The phenotype data "AD Challenge Training Data: Clinical (Updated)" were used in order to maintain consistency with a series of challenges put forth by ADNI in 2014. The AD DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge invited researchers and analysts from all over the world to use statistics and machine learning methods to try to address one of three subchallenges. Subchallenge 3 had the stated goal to "Classify individuals into diagnostic groups using MR imaging." This can be done either by predicting the Mini-Mental State Exam (MMSE) or by using a classification model, such as binomial or multinomial logistic regression models.

The MMSE is a diagnostic tool used to classify adults based on levels of cognitive impairment (Pangman et al., 2000). The exam itself poses a series of questions to test the subject's ability to perform everyday tasks. The maximum score that one can achieve on the test is 30, and the scores take integer values. These scores can then be used in conjunction with other information to diagnose an individual. In both the training and test samples used within this analysis, scores ranged from 20 to 30.

The challenge was judged based on correlations between actual and predicted MMSE scores, which motivated the choice to model with the MMSE as the response variable. More specifically, our analysis used the centered MMSE scores as

the response $\mathbf{y} = (y_1, \ldots, y_n) = (\text{MMSE}_1, \ldots, \text{MMSE}_n) - \overline{\text{MMSE}}$, where $\overline{\text{MMSE}}$ is the average MMSE score. This centering was done in both the training and testing datasets so that an intercept term would not need to be estimated, which could cause identifiability problems with the other non-tensor coefficients in the model. All data were accessed on September 30, 2019. For subjects with multiple scans within the dataset, the scan used was matched to the scan referenced within the ADNI data table found within the csv file labeled `ADNI_Training_Q3_APOE_CollectionADNI1Complete1Yr 1.5T_July22.2014`, distributed on the ADNI website. Some differences between the data used in this analysis and the data used in the challenge must be highlighted. First, the testing data were not publicly released, and our efforts to access the data from the lab that provided them were rebuffed over concerns about privacy. This meant that, in order to have a test dataset, the training dataset from the challenge had to be subset to form the training and test datasets used in this analysis. Second, the contestants also used subject data about their single nucleotide polymorphisms (SNPs), which require their own special methods for processing and model use. As this article focuses on tensor regression rather than other specialized methods, the SNP data are omitted from our modeling process. Third, as our method uses volumetric data, we use the scans themselves, rather than the preprocessed cortical surface data derived from the scans. Cortical surface information is a form of preprocessing that structures the contents of an fMRI scan into a network based on physical proximity accounting for folds and curvature in the brain. Since the transformation functions for changing volumetric data to cortical surface data were not released, we were unable to invert those transformations to obtain the volumetric data that followed the same preprocessing pipeline as the one used in the challenge. This also affected the number of subjects that produced usable

data, which is further explained with the preprocessing steps below.

Brain extraction was performed on the downloaded scans using the `fsl_bet` function from the `fslr` package in R (Muschelli et al., 2015; Smith, 2002a). Finally, affine linear registration to the MNI152 2mm T1-weighted brain template was done using the `flirt` function from the `fslr` package in R (Muschelli et al., 2015; Jenkinson and Smith, 2001; Jenkinson et al., 2002). After these functions were performed, some scans needed to be removed from the sample, as not all of the scans in the dataset were properly labeled to orient the structural scan within the neuroimaging software. The result was a total of 403 usable structural scans that matched to subjects within the ADNI dataset. From these subjects, 303 were randomly assigned to a training dataset, and 100 were assigned to a test dataset. This resulted in a training dataset with 172 males and 131 females, with a mean age of 76.01 years and standard deviation of 6.68 years, mean education level of 15.42 years and standard deviation of 3.04 years, and mean MMSE score of 27.03 and a standard deviation of 2.47. The testing dataset had 60 males and 40 females, with a mean age of 75.49 years and standard deviation of 6.61 years, mean education level of 16.00 years and standard deviation of 2.91 years, and mean MMSE score of 26.65 and standard deviation of 2.84. Since Alzheimer's disease (AD) is found to be significantly associated with larger ventricular volume in the brain (Nestor et al., 2008), the 45th axial slice was chosen for analysis. Since the ventricles are filled with fluid rather than white matter from the brain, a positive association with the MMSE scores in the region of the ventricles would be consistent with the results of Nestor et al. (2008). The selected axial slice contains ventricular volume in healthy adults, which would be expected to expand at a faster rate with age in adults with Alzheimer's disease. The values within each subject's image slice had the mean value for the nonzero tensor covariate

voxels subtracted from them, and then were divided by the standard deviation of that subject's nonzero voxels to remove any subject-specific effect in terms of the unitless measure from the scanner. Finally, the parts of the axial slice that fell outside the brain region within the template were removed from the covariate tensor in order to try to improve the coefficient estimation. This resulted in a tensor covariate for each subject $\mathbf{X}_i \in \mathbb{R}^{70\times 87}$.

For the non-tensor covariates in each subject, $\boldsymbol{\eta}_i$, the number of years of education and the number of Apolipoprotein E4 (APOE4) alleles present in a subject's DNA (0, 1, or 2). The number of years of education shows a significant association with the MMSE scores in an exploratory data analysis. APOE4 has been identified as a genetic risk factor for Alzheimer's disease (Strittmatter and Roses, 1996). The APOE4 covariate was treated as continuous, as exploratory data analysis suggests a linear association with the MMSE.

For the comparison of models in these scenarios, the root mean squared predictive error and the Pearson correlation between the predictions and the actual values in the test dataset are used. The Pearson correlation is included because it was used as a performance metric in the DREAM challenge. Each Bayesian model is run for 11,000 iterations, after which 1,000 iterations are discarded as a burn-in. The point predictions for the response are calculated as the means of the posterior predictive distributions, which are found for each subject as

$$\hat{y}_i^{(s)} = \langle \mathbf{B}^{(s)}, \mathbf{X}_i \rangle + \boldsymbol{\gamma}^{(s)'} \boldsymbol{\eta}_i, \tag{4.3}$$

for each sample s from the posterior distribution. The frequentist tensor regression models are run until the log-likelihood change between steps is less than 0.1, and they predict the response value for each subject using their point estimates for $\mathbf{B}$ and $\boldsymbol{\gamma}$.

Final point estimates for the tensor coefficient are found in the Bayesian models by using the sequential 2-means post-hoc variable selection method (Li and Pati, 2017). The frequentist tensor regression models are not corrected for multiple testing, as they use the LASSO to select the coefficients that are significantly different from zero. The two-step GLM model estimate is found by applying the Benjamini-Hochsberg multiple testing correction and setting voxels that are not significant to have values of zero. The deviance information criterion is used to select which Bayesian model should be used to fit the data, and the Bayesian information criterion is used to select the frequentist rank models. The plots for the final estimates can be seen in Figure 4.5. These estimates are very different among the different models, which may suggest that there is very low signal within these data to conclude that there are nonzero coefficient values. Nonetheless, the point estimate from the BTR Tucker model shows some positive associations along the medial wall and left posterior ventricle, which is consistent with studies that show that increased ventricle volume is associated with Alzheimer's disease (Nestor et al., 2008), and thus, the MMSE. Point estimates for the non-tensor coefficients ($\boldsymbol{\gamma}$) were found as the posterior medians in the Bayesian models after the burn-in. The estimates from the selected competitor models can be seen in table 4.3, along with the 95% posterior credible intervals for the Bayesian models.

In order to further verify the fits of the models, the model fits to the training data are used to predict the MMSE values in the test dataset. The Bayesian models produce predictions as the means of their posterior predictive distributions. The two-step GLM results are used to predict after the Benjamini-Hochsberg multiple testing correction is applied to the tensor coefficient. That is, any voxels within the tensor coefficient that are not deemed to be statistically significant are assigned a value of zero. The frequentist tensor regression models do not

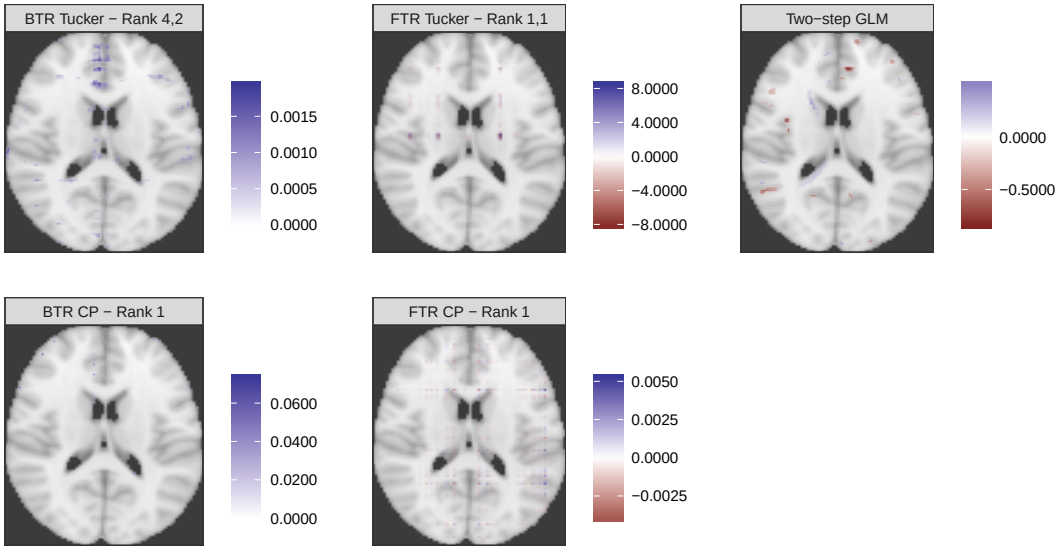

Figure 4.5: Point estimates for the coefficients in the 45th axial slice of the ADNI data training subset.

have any multiple testing correction applied, due to their use of the LASSO. For comparison, the root mean squared predictive error and the Pearson correlation are calculated for each of the selected models. These values can be seen in table 4.3. From the model predictions, the BTR Tucker and FTR CP models perform best, though it is important to note that these two models produce coefficient estimates that are very close to zero, further suggesting that there may be very low signal in the dataset for a significant association between the MMSE and the 45th axial slice of the structural MRI scan.

4.5 Discussion

The Bayesian tensor regression using the Tucker decomposition shows improvement over other tensor regression methods within simulation studies, under the assumption of normally distributed error. Through simulation tests, it is shown

Model	**RMSPE**	**Pearson**
BTR Tucker - Rank 4,2	2.752	0.236
BTR CP - Rank 1	2.772	0.220
FTR Tucker - Rank 1,1	4.438	0.029
FTR CP - Rank 1	2.761	0.215
Two-step GLM	161.308	0.057
No-image	2.762	0.213
Model	$\hat{\gamma}_{\text{Edu}}$ (95% Cred. Int.)	$\hat{\gamma}_{\text{APOE4}}$ (95% Cred. Int.)
BTR Tucker - Rank 4,2	0.135 (-0.709,1.073)	-0.814 (-1.546,0.127)
BTR CP - Rank 1	0.084 (-1.274,1.031)	-0.842 (-1.839,0.126)
FTR Tucker - Rank 1,1	0.074	-0.953
FTR CP - Rank 1	0.04	-0.914
Two-step GLM	0.04	-0.914
No-image	0.04	-0.914

Table 4.3: The root mean squared prediction error, Pearson correlation for predictions in the test data, and the final point estimates of the non-tensor coefficients for the selected competitor models.

that the BTR Tucker models slightly outperform the frequentist Tucker tensor regression models, and it more drastically outperform the Bayesian tensor regression competitor models based on the CP/PARAFAC tensor decomposition. These performance improvements are seen in the inference for both the tensor-valued coefficients, and the vector-valued coefficients.

In the analysis of the ADNI data, it was found that all of the different competing models resulted in an inconclusive final inference about the voxels within the axial slice that show significant correlations with the MMSE scores. Comparing the predictions made with the test data show that the BTR Tucker model outperforms the BTR CP model, if only because it shrinks values in the tensor coefficient to zero.

Future extensions of this model that are planned include using expectation-maximization or variational algorithms to rapidly determine point estimates and select for optimal ranks for a full Markov chain Monte Carlo posterior distribution

simulation. These methods would enhance the analysis while speeding it up considerably. Additionally, formalizing the software used in these and other Bayesian tensor regression methods is a priority, in order to make these analyses accessible to neuroimagers that have not been formally trained in Bayesian statistical modeling.

Chapter 5

Conclusion

Within this dissertation, we have covered some new methods and applications that can be used to incorporate tensor-structured data into Bayesian regression analyses. All of the methods employed tensor decompositions in order to reduce the parameter, shrinking the parameter space while preserving expected spatial structures within the tensor-valued parameter spaces. We began with the development and theoretic validation of a new prior for modeling scenarios with a tensor-valued response and scalar covariates. Expanding on work done in Guhaniyogi et al. (2017), the prior relies on the CP/PARAFAC tensor decomposition (Tucker, 1966) and a stick-breaking structure with proven posterior consistency under mild assumptions. Through testing on simulated datasets, the prior was shown to offer attractive estimation of a sparse tensor-valued coefficient. The prior was then implemented in the analysis of single-subject fMRI data (Schonberg et al., 2012), showing results consistent with beliefs about risk-processing from neuroimaging literature. From there, we expanded the model structure using the CP/PARAFAC tensor decomposition further to apply to scenarios with multiple tensor responses with graphical inference to estimate network structures between the different response tensors. This method was then applied to fMRI data with multiple sub-

jects, providing applicable inference to whole brain and high-resolution brain slice data in identifying risk-processing voxels within the brain while also identifying a connectivity network between different regions of interest within the brain. Finally, we explored the use of the Tucker tensor decomposition (Tucker, 1966) to flexibly regress scalar responses on tensor- and scalar-valued covariates. Simulated data analyses show improvements in inference over other models using tensor decompositions proposed in Zhou et al. (2013), Li et al. (2018), and Guhaniyogi et al. (2017). The model was then used to analyze data including structural magnetic resonance images from the Alzheimer's Disease Neuroimaging Initiative and make inference on parts of the brain associated with scores on cognitive exams used to diagnose Alzheimer's Disease. Results show improved inference and prediction abilities, confirming accepted ideas from the neuroscience community about the structures within the brain that are associated with the disease.

I began this journey into tensor regression with the expectation that I would be developing models to deal with "big data" problems. I came to appreciate the size of the data and the computational challenges they impose through tests that lasted hours and real data analyses that lasted days or even weeks. This research has led to a deep appreciation for parallel computing and the computational resources of large universities. The size of the data and the use of Bayesian methods motivated me to study methods in computational efficiency. I have dabbled in other programming languages (C, Python, Julia, etc.) in order to try and decrease the amount of time I waited to see if my Markov Chain Monte Carlo simulations would work under hundreds of different settings. In the end, I have stuck with R in my software due to its adoption in many research groups, but the lure of greater efficiency may eventually draw me away.

Working with neuroimaging data required learning enough neuroscience to

avoid modeling mistakes that rendered my inference, no matter how sophisticated, completely worthless. I still find myself learning important pieces of information about data that I have spent years looking at. Many people helped me pick up information, which I gathered piecemeal, like breadcrumbs. Andrew Jahn, who is now at the University of Michigan, freely posted many tutorials and resources online that gave me a baseline of understanding using the open-source neuroimaging software FSL. I learned more through the SpaceTime meetings with contributors from UC Irvine, KARST, the University of Minnesota, Rice University, and others. Meeting other neuroimaging analysts at different workshops and conferences changed my understanding of the problems that I was attempting to address through modeling. When I first started earning my PhD, I believed that I wanted the education to attain a high-paying job in the technology industry, but this work has changed my career focus to the modeling of neuroimaging data.

Several expansions on these investigations are being considered as future work to improve on the adoption and inference from these methods. An R software package is in development with the goal of making these Bayesian inferential tools available to neuroimaging labs, complete with data simulation and model diagnostic tools. Formal comparisons of other shrinkage priors on components in tensor decompositions may yield improved results in the inference of the tensor-valued coefficients. Additional ongoing research with Dr. Amanda Mejia at Indiana University is underway to implement a Bayesian method of the analysis of fMRI data. Current work is focused on the software implementation of the methods described in Mejia et al. (2019) to analyze cortical surface fMRI data, with multiple research projects planned in the area of Bayesian neuroimaging analysis. For example, developing Bayesian methods in modeling the haemodynamic response function (HRF) coefficients may significantly improve inference by removing the assump-

tion of equal HRFs for different subjects. For analyses in which point estimates in the tensor-valued parameters are desired, but full posterior distribution inference is unnecessary, further research explorations in expectation-maximization and variational Bayes methods are promising.

Appendix A

Bayesian Tensor Response Regression With an Application to Brain Activation Studies

A.1 Proofs

The proof of Theorem 2.3.1 relies in part on the existence of exponentially consistent sequence of tests.

Definition An exponentially consistent sequence of test functions Φ_T for testing $H_0 : \mathbf{B}_{(T)} = \mathbf{B}_{(T)}^{(0)}$ vs. $H_1 : \mathbf{B}_{(T)} \in \mathcal{A}_T$ satisfies

$$E_{\mathbf{B}_{(T)}^{(0)}}(\Phi_T) \leq c_1 \exp(-b_1 T), \qquad \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} E_{\mathbf{B}_{(T)}}(1 - \Phi_T) \leq c_2 \exp(-b_2 T)$$

for some $c_1, c_2, b_1, b_2 > 0$.

Theorem A.1.1. *There exist an exponentially consistent sequence of tests Φ_T for testing $H_0 : \mathbf{B}_{(T)} = \mathbf{B}_{(T)}^{(0)}$ vs. $H_1 : \mathbf{B}_{(T)} \in \mathcal{A}_T$.*

Proof. Let $\zeta \in \mathcal{F}_1 \times \mathcal{F}_2$. For any $\boldsymbol{h}_1 \in \zeta_1$, let

$$\hat{\mathbf{B}}_{(T),\boldsymbol{h}_1,\zeta_{2,\boldsymbol{h}_1}} = (\boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}_1}} \boldsymbol{R}^{-1} \boldsymbol{X}_{\zeta_{2,\boldsymbol{h}_1}})^{-1} \boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}_1}} \boldsymbol{R}^{-1} \boldsymbol{y}_{\boldsymbol{h}_1},$$

where $\boldsymbol{y}_{\boldsymbol{h}_1} = (Y_{1,\boldsymbol{h}_1}, ..., Y_{T,\boldsymbol{h}_1})'$ and $\boldsymbol{X}_{\zeta_{2,\boldsymbol{h}_1}}$ is a $T \times |\zeta_{2,\boldsymbol{h}_1}|$ dimensional matrix whose tth row is given by $(\boldsymbol{x}_{j,t} : j \in \zeta_{2,\boldsymbol{h}_1})$. Define a test function $\Phi_T = \max_{|\zeta| \leq \tilde{s}_T + s_T, \zeta \supseteq \zeta^{(0)}} 1\left\{||\hat{\mathbf{B}}_{(T),\zeta} - \mathbf{B}^{(0)}_{(T),\zeta}||_2 > \epsilon/4\right\}$. In what follows, we will show that Φ_T is an exponentially consistent sequence of tests.

$$\begin{aligned}
&E_{\mathbf{B}^{(0)}_{(T)}}(\Phi_T) \leq \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(||\hat{\mathbf{B}}_{(T),\zeta} - \mathbf{B}^{(0)}_{(T),\zeta}||_2 > \epsilon/4\right) \\
&\leq \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(\sum_{\boldsymbol{h} \in \zeta_1} \boldsymbol{\Delta}_{\boldsymbol{h}} > \epsilon^2/16\right) \\
&\text{where } \boldsymbol{\Delta}_{\boldsymbol{h}} = (\hat{\mathbf{B}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{B}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}})'(\hat{\mathbf{B}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{B}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}}) \\
&\leq \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\Big(\sum_{\boldsymbol{h} \in \zeta_1} \boldsymbol{\Delta}'_{\boldsymbol{h}} > T\lambda_0^2\epsilon^2/16\Big) \\
&\text{where } \boldsymbol{\Delta}'_{\boldsymbol{h}} = (\hat{\mathbf{B}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{B}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}})'(\boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}}} \boldsymbol{R}^{-1} \boldsymbol{X}_{\zeta_{2,\boldsymbol{h}}})(\hat{\mathbf{B}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{B}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}}) \\
&= \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(\sum_{\boldsymbol{h} \in \zeta_1} \chi^2_{|\zeta_{2,\boldsymbol{h}}|} > T\lambda_0^2\epsilon^2/16\right) \\
&= \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(\chi^2_{|\zeta|} > T\lambda_0^2\epsilon^2/16\right) \leq \binom{p_{(T)}}{\tilde{s}_{(T)} + s_{(T)}} \exp(-T\lambda_0^2\epsilon^2/16),
\end{aligned}$$

where the last inequality follows from Lemma A.1 and A.2 in Song and Liang (2017). Note that

$$\binom{p_{(T)}}{\tilde{s}_{(T)} + s_{(T)}} \leq p_{(T)}^{\tilde{s}_{(T)} + s_{(T)}} \leq \exp((\tilde{s}_{(T)} + s_{(T)}) \log(p_{(T)})) \leq \exp(T\lambda_0^2\epsilon^2/32),$$

by assumptions (b) and (c). Thus $E_{\mathbf{B}^{(0)}_{(T)}}(\Phi_T) \leq \exp(-T\lambda_0^2\epsilon^2/32)$. Let $\tilde{\zeta} = \zeta^{(0)} \cup \{(\boldsymbol{h}_1, h_2) : |\mathbf{B}_{(T),\boldsymbol{h}_1,h_2}| \geq a_T\}$

$$\sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} E_{\mathbf{B}_{(T)}}(1-\Phi_T) \leq \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} E_{\mathbf{B}_{(T)}}(1 - 1\left\{||\hat{\mathbf{B}}_{(T),\tilde{\zeta}} - \mathbf{B}^{(0)}_{(T),\tilde{\zeta}}||_2 > \epsilon/4\right\})$$
$$= \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} P_{\mathbf{B}_{(T)}}\left(||\hat{\mathbf{B}}_{(T),\tilde{\zeta}} - \mathbf{B}^{(0)}_{(T),\tilde{\zeta}}||_2 \leq \epsilon/4\right).$$

Under $\mathcal{A}_T$, $||\mathbf{B}_{(T),\tilde{\zeta}} - \mathbf{B}^{(0)}_{(T),\tilde{\zeta}}||_2 \geq ||\mathbf{B}_{(T)} - \mathbf{B}^{(0)}_{(T)}||_2 - ||\mathbf{B}_{(T),\tilde{\zeta}^c} - \mathbf{B}^{(0)}_{(T),\tilde{\zeta}^c}||_2 \geq \epsilon - a_T p_T \geq \epsilon/2$. Where the last inequality follows due to the fact $\mathbf{B}^{(0)}_{(T),\tilde{\zeta}^c} = \mathbf{0}$ and $|\mathbf{B}_{(T),\boldsymbol{h}_1,h_2}| \leq a_T$ for $(\boldsymbol{h}_1, h_2) \in \tilde{\zeta}^c$.

Using the above fact

$$\sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} E_{\mathbf{B}_{(T)}}(1-\Phi_T) \leq \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} P_{\mathbf{B}_{(T)}}\left(||\hat{\mathbf{B}}_{(T),\tilde{\zeta}} - \mathbf{B}^{(0)}_{(T),\tilde{\zeta}}||_2 \leq \epsilon/4\right)$$
$$\leq \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} P_{\mathbf{B}_{(T)}}\left(||\hat{\mathbf{B}}_{(T),\tilde{\zeta}} - \mathbf{B}_{(T),\tilde{\zeta}}||_2 \geq -||\hat{\mathbf{B}}_{(T),\tilde{\zeta}} - \mathbf{B}^{(0)}_{(T),\tilde{\zeta}}||_2 + ||\mathbf{B}_{(T),\tilde{\zeta}} - \mathbf{B}^{(0)}_{(T),\tilde{\zeta}}||_2\right)$$
$$\leq \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} P_{\mathbf{B}_{(T)}}\left(||\hat{\mathbf{B}}_{(T),\tilde{\zeta}} - \mathbf{B}_{(T),\tilde{\zeta}}||_2 \geq \epsilon/4\right)$$
$$\leq \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} P\left(\sum_{\boldsymbol{h} \in \zeta_1} \boldsymbol{\Delta}_{\boldsymbol{h}} > T\lambda_0^2\epsilon^2/16\right)$$
where $\boldsymbol{\Delta}_{\boldsymbol{h}} = (\hat{\mathbf{B}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{B}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}})'(\boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}}}\boldsymbol{R}^{-1}\boldsymbol{X}_{\zeta_{2,\boldsymbol{h}}})(\hat{\mathbf{B}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{B}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}})$
$$\leq \sup_{\mathbf{B}_{(T)} \in \mathcal{A}_T} P\left(\sum_{\boldsymbol{h} \in \zeta_1} \chi^2_{|\zeta_{2,\boldsymbol{h}}|} > T\lambda_0^2\epsilon^2/16\right)$$
$$\leq P\left(\chi^2_{|\zeta|} > T\lambda_0^2\epsilon^2/16\right) \leq \exp(-T\lambda_0^2\epsilon^2/16).$$

Hence Φ_T is a exponentially consistent sequence of tests. □

Next, we provide a bound on the discrepancy between the true and fitted tensor.

Theorem A.1.2. *Let* $\mathcal{K}(\theta) = -\log\{\Pi_T(\mathbf{B}_{(T)} : ||\mathbf{B}_{(T)} - \mathbf{B}^{(0)}_{(T)}||_\infty < \theta)\}$ *and* $\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} = (\beta_{j,1,k,v_j,(T)}, ..., \beta_{j,R,k,v_j,(T)})'$, *and* $\tilde{\boldsymbol{\beta}}^{(0)}_{j,k,v_j,(T)} = (\beta^{(0)}_{j,1,k,v_j,(T)}, ..., \beta^{(0)}_{j,R,k,v_j,(T)})'$, $\beta^{(0)}_{j,r,k,v_j,(T)} = 0$ *for* $r \in \{R_0 + 1, \ldots, R\}$, $R > R_0$. *For* $k = 1, ..., m_{(T)}$, *assume that* $\Delta_{\boldsymbol{v},k}$ *is a positive root of the equations given, for all* $\boldsymbol{v} \in \mathcal{F}_1 \times \mathcal{F}_2$, *by*

$$
\begin{aligned}
&x(x + ||\tilde{\boldsymbol{\beta}}^{(0)}_{2,k,v_2,(T)}||) \cdots (x + ||\tilde{\boldsymbol{\beta}}^{(0)}_{D,k,v_D,(T)}||) + \\
&||\tilde{\boldsymbol{\beta}}^{(0)}_{1,k,v_1,(T)}||x(x + ||\tilde{\boldsymbol{\beta}}^{(0)}_{2,k,v_2,(T)}||) \cdots (x + ||\tilde{\boldsymbol{\beta}}^{(0)}_{D,k,v_D,(T)}||) + \\
&\cdots + x||\tilde{\boldsymbol{\beta}}^{(0)}_{2,k,v_2,(T)}|| \cdots ||\tilde{\boldsymbol{\beta}}^{(0)}_{D,k,v_D,(T)}|| - \theta = 0, \qquad \text{(A.1)}
\end{aligned}
$$

and $\Delta = \min_{\boldsymbol{v},k} \Delta_{\boldsymbol{v},k}$. *Then, for some constant* C,

$$
\begin{aligned}
\mathcal{K}(\theta) \leq &\left(R m_{(T)} \sum_{j=1}^{D} p_{j,(T)} \right) \ln \left\{ \frac{(2\pi R)^{1/2}}{(2\Delta)} \right\} - \ln(C) + R m_{(T)} \sum_{j=1}^{D} \ln \left\{ \frac{\Gamma(a_\lambda)}{\Gamma(a_\lambda + p_{j,(T)})} \right\} \\
&+ \sum_{k=1}^{m_{(T)}} \sum_{j=1}^{D} \sum_{r=1}^{R_0} (a_\lambda + p_{j,(T)}) \ln \left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}} \{ (\beta^{(0)}_{j,r,k,v_j,(T)})^2 + 2\Delta^2 \}^{1/2} \right] \\
&+ (R - R_0) m_{(T)} \sum_{j=1}^{D} (a_\lambda + p_{j,(T)}) \ln\Big(b_\lambda + p_{j,(T)} 2^{1/2} \Delta \Big).
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&|\mathbf{B}_{\boldsymbol{v},k,(T)} - \mathbf{B}^{(0)}_{\boldsymbol{v},k,(T)}| \\
&= \left| \sum_{r=1}^{R} \beta_{1,r,k,v_1,(T)} \cdots \beta_{D,r,k,v_D,(T)} - \sum_{r=1}^{R} \beta^{(0)}_{1,r,k,v_1,(T)} \cdots \beta^{(0)}_{D,r,k,v_D,(T)} \right| \\
=& \left| \sum_{r=1}^{R} \left\{ (\beta_{1,r,k,v_1,(T)} - \beta^{(0)}_{1,r,k,v_1,(T)}) \prod_{j \neq 1} \beta_{j,r,k,v_j,(T)} + \cdots \right. \right. \\
&\left. \left. + (\beta_{D,r,k,v_D,(T)} - \beta^{(0)}_{D,r,k,v_D,(T)}) \prod_{j \neq D} \beta^{(0)}_{j,r,k,v_j,(T)} \right\} \right| \\
\leq& ||\tilde{\boldsymbol{\beta}}_{1,k,v_1,(T)} - \tilde{\boldsymbol{\beta}}^{(0)}_{1,k,v_1,(T)}||_2 \prod_{j \neq 1} ||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)}||_2 + \cdots \\
&+ ||\tilde{\boldsymbol{\beta}}_{D,k,v_D,(T)} - \tilde{\boldsymbol{\beta}}^{(0)}_{D,k,v_D,(T)}||_2 \prod_{j \neq D} ||\tilde{\boldsymbol{\beta}}^{(0)}_{j,k,v_j,(T)}||_2,
\end{aligned}
$$

Note that (A.1) can be written as $\mathrm{g}_{\boldsymbol{v},k}(x) = 0$, where

$$
\mathrm{g}_{\boldsymbol{v},k}(x) = a_{D,k,\boldsymbol{v}} x^D + \cdots + a_{1,k,\boldsymbol{v}} x - a_{0,k,\boldsymbol{v}}
$$

and the $a_{j,k,\boldsymbol{v}}$'s are suitably chosen to match the coefficient of x^j in (A.1). By Cauchy's bound on the roots of polynomials, Eq. (A.1) has only one positive root, namely the real $\Delta_{\boldsymbol{v},k}$ that satisfies $\Delta_{\boldsymbol{v},k} \leq 1 + \max_{j=0,\ldots,D} |a_{j,k,\boldsymbol{v}}|$, for all $\boldsymbol{v}$ and k. From (A.1), the fact that $||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\beta}}^{(0)}_{j,k,v_j,(T)}|| < \Delta$ for all $v_j \in \{1, \ldots, p_{j,(T)}\}$, $j \in \{1, \ldots, D\}$ and $k \in \{1, ..., m_{(T)}\}$ implies

$$
|\mathbf{B}_{\boldsymbol{v},k,(T)} - \mathbf{B}^{(0)}_{\boldsymbol{v},k,(T)}| \leq \mathrm{g}_{\boldsymbol{v},k}(\Delta) + \theta \leq \mathrm{g}_{\boldsymbol{v},k}(\Delta_{\boldsymbol{v},k}) + \theta = \theta,
$$

which leads to $||\mathbf{B}_{(T)} - \mathbf{B}^{(0)}_{(T)}||_\infty < \theta$. Hence

$$\Pi_T(\mathbf{B}_{(T)} : ||\mathbf{B}_{(T)} - \mathbf{B}_{(T)}^{(0)}||_\infty < \theta)$$
$$\geq \Pi_T(\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}\ ||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)}^{(0)}||_2 < \Delta).$$

We will bound the right-hand side from below.

$$\Pi_T\Big(\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}\ ||\tilde{\boldsymbol{\beta}}_{j,v_j,T} - \tilde{\boldsymbol{\beta}}_{j,v_j,T}^{(0)}||_2 <$$
$$\Delta \mid \forall_{k\in\{1,\ldots,m_{(T)}\}}\ \{\phi_{r,k}\}, \tau_k, \{W_{jr,k}\}\Big)$$
$$= \prod_{k=1}^{m_{(T)}}\prod_{j=1}^{D}\prod_{v_j=1}^{p_{j,(T)}}\left[\exp\left\{-\sum_{r=1}^{R}\frac{\beta_{j,r,k,v_j,(T)}^{(0)})^2}{2w_{j,r,k,v_j}\phi_{r,k}\tau_k}\right\}\times\right.$$
$$\left.\Pi_T\left(||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)}|| < \frac{\Delta}{2} \mid \{\phi_{r,k}\}, \tau_k, \{W_{jr,k}\}\right)\right]$$
$$\geq \prod_{k=1}^{m_{(T)}}\prod_{j=1}^{D}\prod_{v_j=1}^{p_{j,(T)}}\left[\exp\left\{-\sum_{r=1}^{R}\frac{\beta_{j,r,k,v_j,(T)}^{(0)})^2}{(2w_{j,r,k,v_j}\phi_{r,k}\tau_k}\right\}\times\right.$$
$$\left.\prod_{r=1}^{R}\left[\exp\left\{\frac{-\Delta^2}{(\phi_{r,k}\tau_k w_{j,r,k,v_j})}\right\}\frac{2\Delta}{(2\pi R\phi_{r,k}\tau_k w_{j,r,k,v_j})^{1/2}}\right]\right]$$
$$\geq \prod_{k=1}^{m_{(T)}}\prod_{j=1}^{D}\prod_{v_j=1}^{p_{j,(T)}}\prod_{r=1}^{R}\left[\frac{2\Delta}{(2\pi R\phi_{r,k}\tau_k w_{j,r,k,v_j})^{1/2}}\exp\left[-\frac{2\Delta^2 + (\beta_{j,r,k,v_j,(T)}^{(0)})^2}{2\phi_{r,k}\tau_k w_{j,r,k,v_j}}\right]\right],$$

where Step 2 follows from Anderson's lemma. Integrating out the w_{j,r,k,v_j}'s, we obtain, $\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}$,

$$\Pi\left(||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)}^{(0)}|| < \Delta \mid \tau_k, \{\phi_{r,k}\}, \{\lambda_{j,r,k}\}\right)$$
$$\geq \prod_{k=1}^{m_{(T)}}\prod_{r=1}^{R}\prod_{j=1}^{D}\left[\left(\frac{2\Delta\lambda_{j,r,k}}{(R\phi_{r,k}\tau_k)^{1/2}}\right)^{p_{j,(T)}}\exp\left[-\lambda_{j,r,k}\sum_{v_j=1}^{p_{j,(T)}}\frac{\left((\beta_{j,r,k,v_j,(T)}^{(0)})^2 + 2\Delta^2\right)^{1/2}}{(\phi_{r,k}\tau_k)^{1/2}}\right]\right].$$

Integrating out the $\lambda_{j,r,k}$'s, we then get, $\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}$,

$$
\begin{aligned}
&\Pi_T\left(||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\beta}}^{(0)}_{j,k,v_j,(T)}|| < \Delta \mid \tau_k, \{\phi_{r,k}\}\right) \\
&\geq \prod_{k=1}^{m_{(T)}}\prod_{r=1}^{R}\prod_{j=1}^{D}\left[\left\{\frac{2\Delta}{(R\phi_{r,k}\tau_k)^{1/2}}\right\}^{p_{j,(T)}} \times \right. \\
&\qquad \left. \frac{\Gamma(a_\lambda + p_{j,(T)})}{\left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}}\{(\beta^{(0)}_{j,r,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2}(\phi_{r,k}\tau_k)^{-1/2}\right]^{a_\lambda + p_{j,(T)}}}\right]\left\{\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\right\}^{R(D)} \\
&\geq \prod_{k=1}^{m_{(T)}}\prod_{r=1}^{R}\prod_{j=1}^{D}\left[\left\{\frac{2\Delta}{(R\phi_{r,k}\tau_k)^{1/2}}\right\}^{p_{j,(T)}}\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)} \times \right. \\
&\qquad \left. \frac{\Gamma(a_\lambda + p_{j,(T)})(\phi_{r,k}\tau_k)^{(a_\lambda + p_{j,(T)})/2}\mathbf{1}\{\tau_k \in (0,1)\}}{\left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}}\{(\beta^{(0)}_{j,r,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2}\right]^{a_\lambda + p_{j,(T)}}}\right]
\end{aligned}
$$

Integrating our $\phi_{r,k}$'s together we obtain, $\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}$,

$$
\begin{aligned}
&\Pi_T\left(||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\beta}}^{(0)}_{j,k,v_j,(T)}|| < \Delta \mid \tau_k\right) \\
&\geq \prod_{k=1}^{m_{(T)}}\prod_{r=1}^{R}\prod_{j=1}^{D}\left[\left\{\frac{2\Delta}{(R\tau_k)^{1/2}}\right\}^{p_{j,(T)}}\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)} \times \right. \\
&\qquad \left. \frac{\Gamma(a_\lambda + p_{j,(T)})\tau_k^{(a_\lambda + p_{j,(T)})/2}\mathbf{1}\{\tau_k \in (0,1)\}}{\left[b_\lambda + \sum_{i_j=1}^{p_{j,(T)}}\{(\beta^{(0)}_{j,r,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2}\right]^{a_\lambda + p_{j,(T)}}}\right] \times \\
&\qquad \prod_{r=1}^{R-1}\left[\frac{Beta(D, \alpha_k + D(R-r))}{Beta(1,\alpha_k)}\right],
\end{aligned}
$$

where $Beta(m_1, m_2)$ is the integrating constant for the Beta density with parameters m_1 and m_2. Finally, integrating out τ_k, leads to,

$\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}$,

$$
\begin{aligned}
&\Pi_T(||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\beta}}^{(0)}_{j,k,v_j,(T)}|| < \Delta) \\
&\geq \prod_{k=1}^{m_{(T)}} \prod_{j=1}^{D} \left\{ \frac{\Gamma(a_\lambda + p_{j,(T)})}{\Gamma(a_\lambda)} \right\}^R \times \\
&\quad \prod_{k=1}^{m_{(T)}} \prod_{j=1}^{D} \prod_{r=1}^{R} \left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}} \{(\beta^{(0)}_{j,r,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2} \right]^{-a_\lambda - p_{j,(T)}} \times \\
&\quad \{2\Delta/(2\pi R)^{1/2}\}^{Rm_{(T)}\sum_{j=1}^{D} p_{j,(T)}} C^{-1},
\end{aligned}
$$

for some constant C. Hence, $\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}$,

$$
\begin{aligned}
\mathcal{K}(\theta) &\leq -\log\left[\Pi_T(||\tilde{\boldsymbol{\beta}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\beta}}^{(0)}_{j,k,v_j,(T)}|| < \Delta)\right] \\
&\leq \left(Rm_{(T)} \sum_{j=1}^{D} p_{j,(T)} \right) \ln\left\{ \frac{(2\pi R)^{1/2}}{(2\Delta)} \right\} - \ln(C) + Rm_{(T)} \sum_{j=1}^{D} \ln\left\{ \frac{\Gamma(a_\lambda)}{\Gamma(a_\lambda + p_{j,(T)})} \right\} \\
&\quad + \sum_{k=1}^{m_{(T)}} \sum_{j=1}^{D} \sum_{r=1}^{R_0} (a_\lambda + p_{j,(T)}) \ln\left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}} \{(\beta^{(0)}_{j,r,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2} \right] \\
&\quad + (R - R_0) m_{(T)} \sum_{j=1}^{D} (a_\lambda + p_{j,(T)}) \ln\Big(b_\lambda + p_{j,(T)} 2^{1/2} \Delta \Big).
\end{aligned}
$$

□

Under assumptions (a)-(f), the R.H.S is $o(T)$. Thus, we present the next theorem whose proof follows immediately from Theorem A.1.2.

Theorem A.1.3. *For any constant $\theta > 0$, under conditions (a)-(f) of Theorem 2.3.1, $\mathcal{K}(\theta) = o(T)$.*

Proof of Theorem 2.3.1

Proof.

$$\Pi_T(\mathcal{A}_T) = \frac{\int_{\mathcal{A}_T} f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})\pi_T(\mathbf{B}_{(T)})}{\int f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})\pi_T(\mathbf{B}_{(T)})} = \frac{\int_{\mathcal{A}_T} \frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}^{(0)}_{(T)})}\pi_T(\mathbf{B}_{(T)})}{\int \frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}^{(0)}_{(T)})}\pi_T(\mathbf{B}_{(T)})}$$
$$= \frac{\mathcal{N}_T}{\mathcal{D}_T} \leq \Phi_T + (1-\Phi_T)\frac{\mathcal{N}_T}{\mathcal{D}_T}, \tag{A.2}$$

where Φ_T is the exponentially consistent sequence of tests given by Lemma A.1.1. Note that

$$P_{\mathbf{B}^{(0)}_{(T)}}\left(\Phi_T > \exp(-T\lambda_0^2\epsilon^2/64)\right) \leq E_{\mathbf{B}^{(0)}_{(T)}}(\Phi_T)\exp(T\lambda_0^2\epsilon^2/64) \leq \exp(-T\lambda_0^2\epsilon^2/64).$$

Therefore $\sum_{T=1}^{\infty} P_{\mathbf{B}^{(0)}_{(T)}}(\Phi_T > \exp(-T\lambda_0^2\epsilon^2/64)) < \infty$. Applying Borel-Cantelli lemma $P_{\mathbf{B}^{(0)}_{(T)}}(\Phi_T > \exp(-T\lambda_0^2\epsilon^2/64)$ infinitely often$) = 0$. Thus,

$$\Phi_T \to 0 \quad a.s. \tag{A.3}$$

In addition, we have

$$E_{\mathbf{B}^{(0)}_{(T)}}((1-\Phi_T)\mathcal{N}_T) = \int(1-\Phi_T)\int_{\mathcal{A}_T}\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}^{(0)}_{(T)})}\pi_T(\mathbf{B}_{(T)})f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}^{(0)}_{(T)})$$
$$= \int_{\mathcal{A}_T}\int(1-\Phi_T)f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})\pi_T(\mathbf{B}_{(T)})$$
$$\leq \sup_{\mathbf{B}_{(T)}\in\mathcal{A}_T} E_{\mathbf{B}_{(T)}}(1-\Phi_T)$$
$$\leq \exp(-T\lambda_0^2\epsilon^2/16).$$

Applying Borel-Cantelli lemma,

$P_{\mathbf{B}_{(T)}^{(0)}}\left((1-\Phi_T)\mathcal{N}_T \exp(T\lambda_0^2\epsilon^2/32) > \exp(-T\lambda_0^2\epsilon^2/64) \text{ infinitely often}\right) = 0$ so

$$\exp(T\lambda_0^2\epsilon^2/32)(1-\Phi_T)\mathcal{N}_T \to 0 \quad a.s.. \tag{A.4}$$

Note that $\mathcal{D}_T = \int \frac{f(\boldsymbol{Y}_1..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)}^{(0)})}\pi_T(\mathbf{B}_{(T)})$. Let $\tilde{b} = \lambda_0^2\epsilon^2/32$. Consider the set $\mathcal{H}_T = \left\{\mathbf{B}_{(T)} : \frac{1}{T}\log\left[\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)}^{(0)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}\right] < \upsilon\right\}$, for $\upsilon = \tilde{b}/2$.

$$\begin{aligned}\exp(\tilde{b}T)\mathcal{D}_T &\geq \exp(\tilde{b}T)\int\limits_{\mathcal{H}_T} \exp\left(-T\frac{1}{T}\log\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)}^{(0)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}\right)\pi_T(\mathbf{B}_{(T)}) \\ &\geq \exp((\tilde{b}-\tilde{b}/2)T)\Pi_T(\mathcal{H}_T).\end{aligned}$$

In view of (A.2), (A.3) and (A.4), it is enough to show that $-\log(\Pi_T(\mathcal{H}_T)) \leq T\tilde{b}/8$.

$$\begin{aligned}&\frac{1}{T}\log\left[\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)}^{(0)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}\right] \\ &\quad= \frac{1}{T}\left[-\frac{1}{2}\sum_{\boldsymbol{v}}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{B}_{\boldsymbol{v},k,(T)}^{(0)}\boldsymbol{x}_k)'\boldsymbol{R}^{-1}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{B}_{\boldsymbol{v},k,(T)}^{(0)}\boldsymbol{x}_k) + \right. \\ &\qquad\left.\frac{1}{2}\sum_{\boldsymbol{v}}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{B}_{\boldsymbol{v},k,(T)}\boldsymbol{x}_k)'\boldsymbol{R}^{-1}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{B}_{\boldsymbol{v},k,(T)}\boldsymbol{x}_k)\right].\end{aligned}$$

$$\begin{aligned}&\Pi_T\Big(\mathbf{B}_{(T)} : \frac{1}{T}\log\left[\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)}^{(0)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{B}_{(T)})}\right] < \upsilon\Big) \\ &\geq \Pi_T\Big(\mathbf{B}_{(T)} : |\frac{1}{2T}\sum_{\boldsymbol{v}}\sum_{k=1}^{m_{(T)}}(\mathbf{B}_{\boldsymbol{v},k,(T)} - \mathbf{B}_{\boldsymbol{v},k,(T)}^{(0)})'\boldsymbol{x}_k'\boldsymbol{R}^{-1}\boldsymbol{x}_k(\mathbf{B}_{\boldsymbol{v},k,(T)} - \mathbf{B}_{\boldsymbol{v},k,(T)}^{(0)})| < \upsilon\Big) \\ &\geq \Pi_T\Big(\mathbf{B}_{(T)} : ||\mathbf{B}_{(T)} - \mathbf{B}_{(T)}^{(0)}||_2^2 < 2\upsilon/\lambda_1^2\Big) \\ &\geq \Pi_T\Big(\mathbf{B}_{(T)} : ||\mathbf{B}_{(T)} - \mathbf{B}_{(T)}^{(0)}||_\infty < \sqrt{2\upsilon/\lambda_1^2}\Big) \geq \exp(-T\tilde{b}/8),\end{aligned}$$

where the third line follows from assumption (e) of Theorem 3.1 and last inequality is immediate by applying Theorem A.1.3. □

A.2 Posterior Sampling Algorithm

This posterior sampling algorithm can be done efficiently by sampling the tensor-covariate-specific variables in parallel. The index $k = 1, \ldots, m$ corresponds to the k^{th} tensor covariate.

(1) Draw α_k via a Griddy-Gibbs algorithm as follows:

   (a) For each possible value of α_k, draw a sample of size $\mathcal{M}$ from the posterior distributions of $\phi_{r,k}$ and τ_k.

   (b) Evaluate the prior density using each of these individual samples using the previous iteration values for all other parameters.

   (c) Average these densities together for each possible value for α_k in the grid, and then sample one value using the averaged densities as weights.

(2) Using the posterior full conditional kernel of

$$
\begin{aligned}
p(\xi_{r,k}|\boldsymbol{\beta}_{\cdot,r,k}, \mathbf{W}_{\cdot,r,k}, \xi_{-r,k}, \tau_k) \propto \\
\xi_{r,k}^{-\sum p_j/2}(1-\xi_{r,k})^{-(R-r)\sum p_j/2} \times \\
\exp\left\{-\frac{1}{\tau_k}\left[\frac{1}{\xi_{r,k}}\sum_{j=1}^{D}\left(\boldsymbol{\beta}_{j,r,k}^T \mathbf{W}_{j,r,k}^{-1}\boldsymbol{\beta}_{j,r,k}\right) + \right.\right. \\
\left.\left.\sum_{h=r+1}^{R}\frac{1}{\xi_{k,h}\prod_{g=r}^{h-1}(1-\xi_g)}\sum_{j=1}^{D}\left(\boldsymbol{\beta}_{j,r,k}^T \mathbf{W}_{j,r,k}^{-1}\boldsymbol{\beta}_{j,r,k}\right)\right]\right\},
\end{aligned}
$$

draw $\xi_{r,k}^{(s)}$ for sample s using a Metropolis-Hastings step with a normal proposal distribution with mean $\xi_{r,k}^{(s-1)}$ and variance 0.01^2. The value for the variance

was chosen such that the datasets tested showed decent mixing. After drawing $\xi_{1,k}, \ldots, \xi_{R-1,k}$, set $\phi_{r,k} = \xi_{r,k} \times \prod_{h=1}^{r-1}(1 - \xi_{h,k})$, and set $\phi_{R,k} = 1 - \sum_{r=1}^{R-1} \phi_{r,k}$.

(3) Draw each τ_k from a generalized inverse Gaussian distribution, $gIG(\nu, \chi, \psi)$, where

$$\nu = a_\tau - \frac{R \sum p_j}{2}, \qquad \chi = \sum_{r=1}^{R} \frac{1}{\phi_{r,k}} \left(\sum_{j=1}^{D} \boldsymbol{\beta}'_{j,r,k} \mathbf{W}^{-1}_{j,r,k} \boldsymbol{\beta}_{j,r,k} \right), \qquad \psi = 2b_\tau$$

(4) Draw each $\lambda_{j,r,k}$ from a

$$\text{Gamma}\left(a_\lambda + p_j, b_\lambda + \frac{1}{\sqrt{\phi_{r,k}\tau_k}} \sum_{\ell=1}^{p_j} |\beta_{j,r,k,\ell}| \right)$$

(5) Draw each $w_{j,r,k,\ell}$ from a generalized Inverse Gaussian distribution, $gIG(\nu, \chi, \psi)$, where

$$\nu = \frac{1}{2} \qquad \chi = \frac{\beta^2_{j,r,k,\ell}}{\tau_k \phi_{r,k}} \qquad \psi = \lambda^2_{j,r,k}$$

(6) When $D = 2$, draw each $\beta_{j,r,k,\ell}$ from a normal distribution with variance

$$\boldsymbol{\Lambda} = \left(\frac{1}{\phi_{r,k}\tau_k w_{j,r,k,\ell}} + \frac{n \sum x_t^2 \boldsymbol{\beta}'_{-j,r,k} \boldsymbol{\beta}_{-j,r,k}}{\sigma_y^2} \right)^{-1}$$

and mean

$$\boldsymbol{\mu} = \boldsymbol{\Lambda} \frac{\sum x_t \boldsymbol{\beta}'_{-j,r,k} \hat{\mathbf{y}}_t}{\sigma_y^2}$$

where

$$\hat{\mathbf{y}}_t = \mathbf{y}_t - x_t \sum_{r' \neq r} \boldsymbol{\beta}_{1,r',k} \circ \boldsymbol{\beta}_{2,r',k}$$

(7) Draw κ from a truncated normal distribution with a lower bound -1, an upper bound of 1, a variance of

$$\frac{\sigma_\epsilon^2}{\sum_{t=2}^{T}\sum_{r'}\epsilon_{t-1,r'}^2},$$

and mean of

$$\frac{\sum_{t=2}^{T}\sum_{r'}\epsilon_{t,r'}\epsilon_{t-1,r'}}{\sum_{t=2}^{T}\sum_{r'}\epsilon_{t-1,r'}^2},$$

where

$\epsilon_{t,r'} = y_{t,l} - \mathbf{B}_{r'}\mathbf{x}_t$.

(8) Draw σ_ϵ^2 from an inverse gamma distribution with shape parameter

$$a_\epsilon + \frac{\prod p_j}{2},$$

and scale parameter

$$b_\epsilon + \frac{1}{2}\sum_{t=2}^{T}\sum_{r'}(\epsilon_{t,r'} - \kappa\epsilon_{t-1,r'})^2$$

Following this algorithm, the MCMC converges rapidly to the region of the maximum likelihood estimator.

Appendix B

Joint Bayesian Estimation of Voxel Activation and Interregional Connectivity in fMRI

B.1 Algorithm for Drawing from the Joint Posterior Distribution

This posterior sampling algorithm can be done efficiently by sampling the region-specific variables in parallel.

At MCMC iteration s:

(1) Draw α_g via a Griddy-Gibbs algorithm as follows:

(a) For each possible value of α_g, draw a sample of size $\mathcal{M}$ from the posterior distributions of $\phi_{g,r}$ and τ_g.

(b) Evaluate the posterior density using each of these individual samples using the $(s-1)$th values for all other parameters.

(c) Average these densities together for each possible value for α_g in the grid, and then sample one value using the averaged densities as weights.

(2) Using the posterior full conditional kernel of

$$
\begin{aligned}
p(\xi_{g,r}|\boldsymbol{\beta}_{g,\cdot,r}, \mathbf{W}_{g,\cdot,r}, \xi_{g,-r}, \tau_g) \propto \\
&\xi_{g,r}^{-\sum p_j/2}(1-\xi_{g,r})^{-(R-r)\sum p_j/2} \\
&\times \exp\left\{-\frac{1}{\tau_g}\left[\frac{1}{\xi_{g,r}}\sum_{j=1}^{D}\left(\boldsymbol{\beta}'_{g,j,r}\mathbf{W}^{-1}_{g,j,r}\boldsymbol{\beta}_{g,j,r}\right)\right.\right. \\
&\left.\left.+\sum_{k=r+1}^{R}\frac{1}{\xi_{g,k}\prod_{\ell=r}^{k-1}(1-\xi_{g,\ell})}\sum_{j=1}^{D}\left(\boldsymbol{\beta}'_{g,j,r}\mathbf{W}^{-1}_{g,j,r}\boldsymbol{\beta}_{g,j,r}\right)\right]\right\},
\end{aligned}
$$

draw $\xi_{g,r}^{(s)}$ for sample s using a Metropolis-Hastings step with a normal proposal distribution with mean $\xi_{g,r}^{(s-1)}$ and variance 0.01^2. The value for the variance was chosen such that movement within the posterior distribution of the individual parameters $\xi_{g,r}$ could not exchange values at each iteration, which better preserves identifiability. After drawing $\xi_{g,1}, \ldots, \xi_{g,R-1}$, set $\phi_{g,r} = \xi_{g,r} \times \prod_{k=1}^{r-1}(1-\xi_{g,k})$, and set $\phi_{g,R} = 1 - \sum_{r=1}^{R-1}\phi_{g,r}$.

(3) Draw each τ_g from a generalized inverse Gaussian distribution, $gIG(\nu, \chi, \psi)$, where

$$
\nu = a_\tau - \frac{R\sum_{j=1}^{D} p_{g,j}}{2}, \quad \chi = \sum_{r=1}^{R}\frac{1}{\phi_{g,r}}\left(\sum_{j=1}^{D}\boldsymbol{\beta}'_{g,j,r}\mathbf{W}^{-1}_{g,j,r}\boldsymbol{\beta}_{g,j,r}\right), \quad \psi = 2b_\tau
$$

(4) Draw each $\lambda_{g,j,r}$ from a

$$\text{Gamma}\left(a_\lambda + p_{g,j}, b_\lambda + \frac{1}{\sqrt{\phi_{g,r}\tau_g}}\sum_{\ell=1}^{p_{g,j}}|\beta_{g,j,r,\ell}|\right)$$

(5) Draw each $\omega_{g,j,r,\ell} \in \mathbf{W}_{g,j,r}$ (remember that $\mathbf{W}_{g,j,r}$ is a diagonal matrix) from a generalized Inverse Gaussian distribution, $gIG(\nu, \chi, \psi)$, where

$$\nu = \frac{1}{2} \qquad \chi = \frac{\beta_{g,j,r,\ell}^2}{\tau_g \phi_{g,r}} \qquad \psi = \lambda_{g,j,r}^2$$

(6) Draw each $\boldsymbol{\beta}_{g,j,r,\ell}$ from a normal distribution with variance

$$\boldsymbol{\Lambda} = \left(\frac{1}{\phi_{g,r}\tau_g}\mathbf{W}_{g,j,r}^{-1} + \text{diag}\left(\frac{\sum_i\sum_t x_{i,t}^2 \text{vec}\mathbf{B}_{g,-j}^2}{\sigma_y^2}\right)\right)^{-1}$$

and mean

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}\frac{\sum_i\sum_t x_{i,t}\mathbf{B}_{g,-j}\left(\tilde{\mathbf{Y}}_{g,i,t}\right)'_{(j)}}{\sigma_y^2}$$

where

$$\tilde{\mathbf{Y}}_{g,i,t} = \mathbf{Y}_{g,i,t} - d_{g,i}\mathbf{1} - x_{i,t}\sum_{\ell\neq r}\boldsymbol{\beta}_{g,1,\ell}\circ\cdots\circ\boldsymbol{\beta}_{g,D,\ell}$$

and

$$\mathbf{B}_{g,-j} = \sum_{r=1}^{R}\boldsymbol{\beta}_{g,1,r}\circ\cdots\circ\boldsymbol{\beta}_{g,j-1,r}\circ\boldsymbol{\beta}_{g,j+1,r}\circ\cdots\circ\boldsymbol{\beta}_{g,D,r}$$

and $(\bullet)_{(j)}$ is the mode-j matricization of $\bullet$.

(7) Draw each $\mathbf{d}_i$ from a normal distribution

$$\begin{pmatrix} d_{1,i} \\ \vdots \\ d_{G,i} \end{pmatrix} = \mathbf{d}_i \sim \mathrm{N}(\boldsymbol{\theta}_i, \mathbf{M})$$

where

$$\mathbf{M} = \left(\boldsymbol{\Sigma} + \frac{T\mathbf{V}}{\sigma_y^2} \right)^{-1}$$

and

$$\boldsymbol{\theta}_i = \mathbf{M} \left(\frac{\sum_\ell \sum_t \tilde{\mathbf{y}}_{i,t,\ell}}{\sigma_y^2} \right)$$

for voxel ℓ, subject i, and time t. T is the number of time steps in the fMRI scan and $\mathbf{V}$ is a diagonal matrix where $\mathbf{V}_{ii}$ is equal to the number of voxels in region i, and

$$\tilde{\mathbf{y}}_{i,t,\ell} = \begin{pmatrix} \tilde{y}_{1,i,t,\ell} \\ \vdots \\ \tilde{y}_{G,i,t,\ell} \end{pmatrix}, \quad \tilde{y}_{g,i,t,\ell} = y_{g,i,t,\ell} - B_{g,\ell} x_{i,t}$$

(8) For each region g, draw δ_g from a gamma $\left(\frac{n}{2} + 1, \frac{S_{gg}+\zeta}{2} \right)$

(9) Draw $\boldsymbol{\eta}$ from a multivariate normal distribution with covariance

$$\boldsymbol{\varphi} = \left((S_{gg} + \zeta) \boldsymbol{\Sigma}_{-g,-g}^{-1} + \mathrm{diag}(1/\boldsymbol{\Upsilon}_{-g,g}) \right)^{-1}$$

and mean

$$\mathbf{L} = -\boldsymbol{\varphi} \mathbf{S}_{g,-g}$$

Set $\boldsymbol{\Sigma}_{g,-g} = \boldsymbol{\Sigma}_{-g,g} = \boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_{g,g} = \delta_g + \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{-g,-g}^{-1} \boldsymbol{\eta}$

(10) For $i > g$, draw $u_{g,i}$ from an inverse Gaussian distribution with mean $\sqrt{\left(\frac{\zeta^2}{\boldsymbol{\Sigma}_{g,i}}\right)}$ and shape ζ^2. Set $\boldsymbol{\Upsilon}_{g,i} = \boldsymbol{\Upsilon}_{i,g} = 1/u_{g,i}$.

(11) Draw ζ from a gamma(a, b) distribution where

$$a = a_\zeta + \frac{G(G+1)}{2} \quad b = b_\zeta + \frac{\sum_i \sum_j |\Sigma_{ij}|}{2}$$

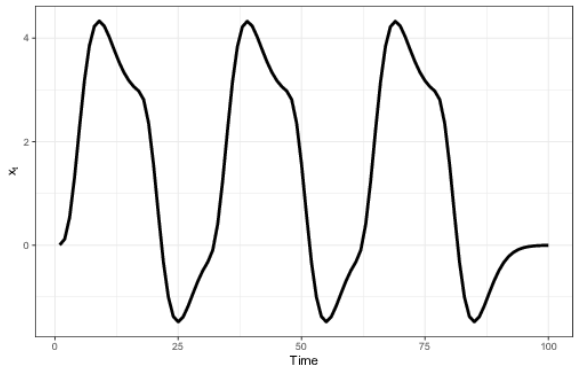

Figure B.1: Values for the covariate x_t in the simulated data.

Convergence Diagnostic Plots Figure B.4 demonstrates convergence through the plot of the log-likelihood across the different ranks in the whole-brain analysis. Figure B.5 shows the autocorrelation function for the different rank models for a randomly-selected voxel within each region of interest.

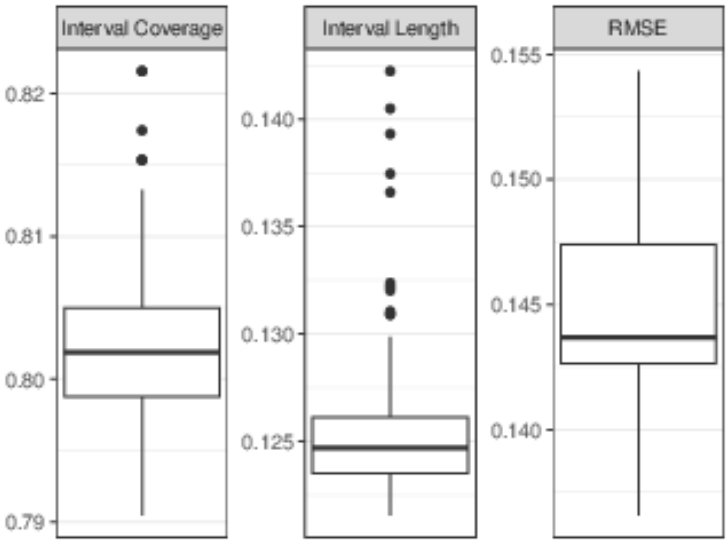

Figure B.2: Boxplots for the 95% interval coverage, interval length, and square root of the mean squared error for the 100 randomly selected hyperparameter settings.

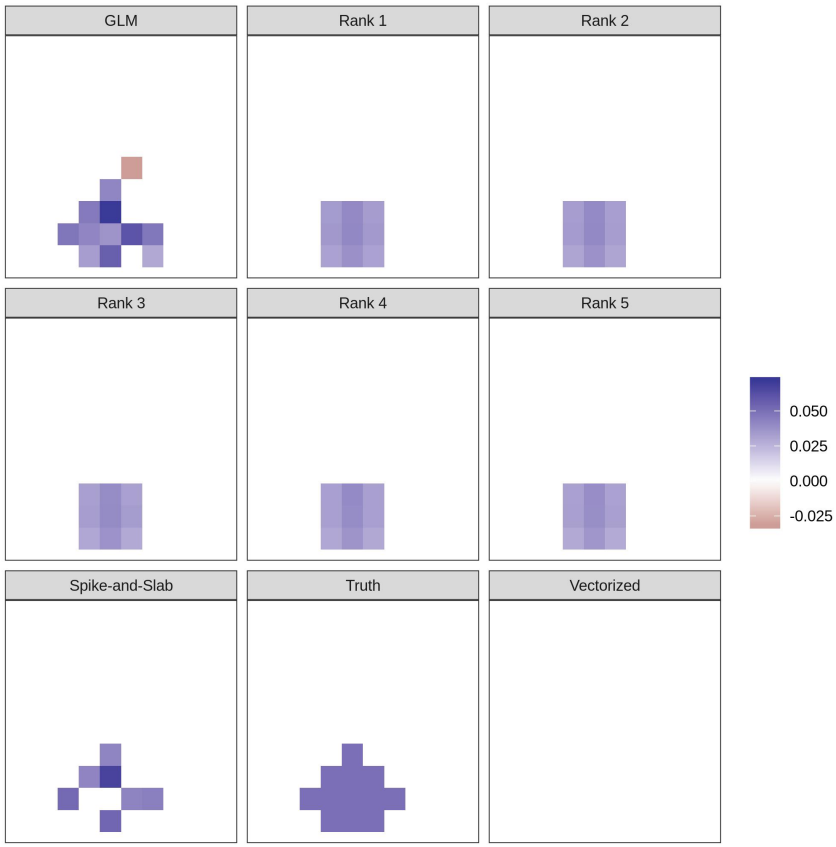

Figure B.3: Rank model estimates and true value for a single slice of a three-dimensional coefficient tensor. Voxels with 99% credible intervals containing zero were set equal to zero. The spike-and-slab and vectorized model estimates are also included for comparison.

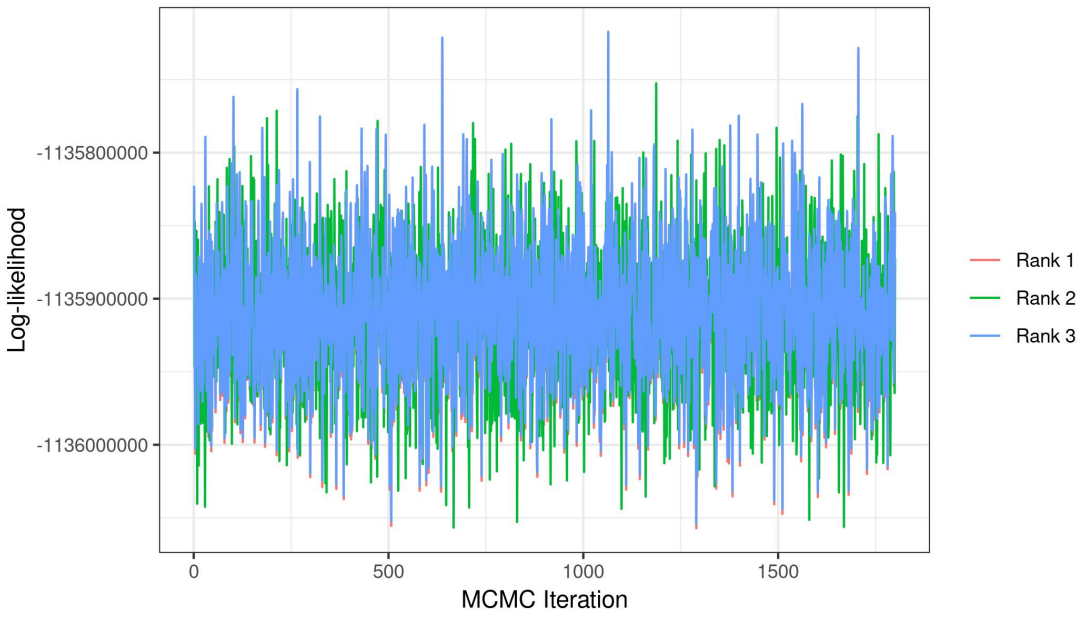

Figure B.4: Log-likelihoods for the Whole Brain Analysis

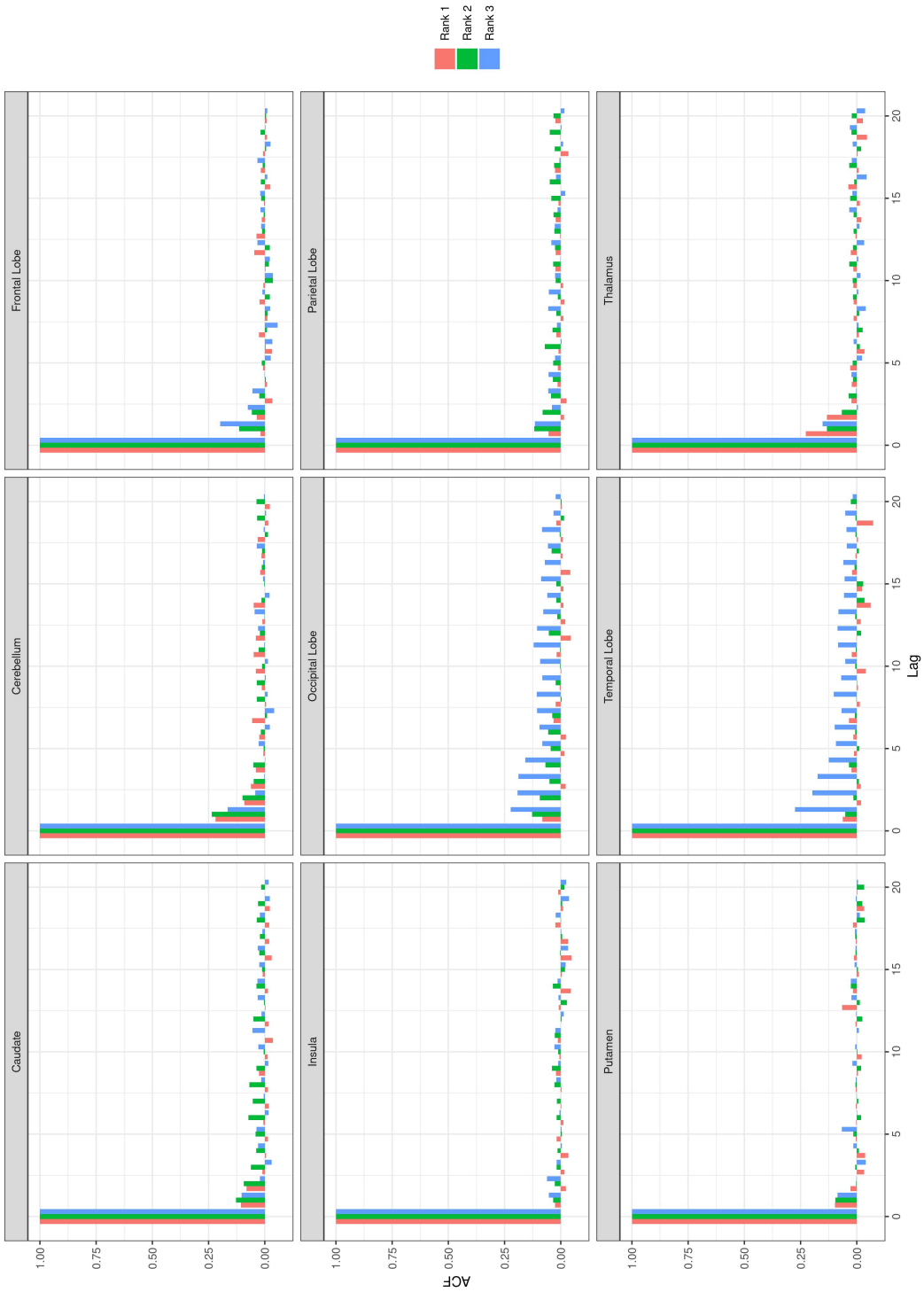

Figure B.5: Autocorrelation Functions for B

Appendix C

Bayesian Tensor Regression Using The Tucker Tensor Decomposition

Posterior Full Conditional Distributions

$\underline{\sigma_y^2}$

$$\sigma_y^2|\mathbf{y},\mathbf{B},\mathbf{X},\boldsymbol{\gamma},\boldsymbol{\eta} \sim \text{Inverse Gamma}\left(a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \langle \mathbf{B}, \mathbf{X}_i\rangle - \boldsymbol{\gamma}'\boldsymbol{\eta}_i\right)^2\right)$$

$\underline{\boldsymbol{\gamma}}$

First, set $\tilde{y}_i = y_i - \langle \mathbf{B}, \mathbf{X}_i\rangle$.

$$\boldsymbol{\gamma}|\mathbf{y},\mathbf{B},\mathbf{X},\boldsymbol{\eta},\sigma_y^2,\mathbf{V} \sim \text{N}\left(\left(\mathbf{V}^{-1} + \frac{1}{\sigma_y^2}\mathbf{I}\right)^{-1}\frac{\boldsymbol{\eta}'\tilde{\mathbf{y}}}{\sigma_y^2}, \left(\mathbf{V}^{-1} + \frac{1}{\sigma_y^2}\mathbf{I}\right)^{-1}\right)$$

$\underline{\boldsymbol{\beta}_{j,r_j}}$

First, call $\tilde{\mathbf{B}} = \mathbf{B} \setminus \boldsymbol{\beta}_{j,r_j}$, and $\mathbf{B}^* = \mathbf{B} \ni \boldsymbol{\beta}_{j,r_j}$. Then,

$$y_i = \langle \mathbf{B}^* + \tilde{\mathbf{B}}, \mathbf{X}_i \rangle + \boldsymbol{\gamma}' \boldsymbol{\eta}_i + \epsilon_i.$$

Next, set

$$\tilde{y}_i = y_i - \boldsymbol{\gamma}' \boldsymbol{\eta}_i - \langle \tilde{\mathbf{B}}, \mathbf{X}_i \rangle.$$

Define the *mode-k matricization* $\mathbf{X}_{(j)}$ of an array $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$ to be $\mathbf{X}_{(j)} \in \mathbb{R}^{p_j \times p_1 p_2 \cdots p_{j-1} p_{j+1} \cdots p_D}$. Set the $\mathbf{B}^*_{-j}$ as the tensor composition of all $\boldsymbol{\beta}_{\ell, r_\ell}$ such that $\ell \in \{1, 2, ..., D\} \setminus j$, which has dimension $D - 1$. Finally using the following notation:

$$\tilde{\mathbf{y}} = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_N \end{pmatrix}, \qquad \mathbf{X}^* = \begin{pmatrix} \mathbf{X}_{1(j)} \mathbf{B}^*_{-j} \\ \vdots \\ \mathbf{X}_{N(j)} \mathbf{B}^*_{-j} \end{pmatrix},$$

the posterior full conditional distribution for $\boldsymbol{\beta}_{j, r_j}$ can be written as

$$\boldsymbol{\beta}_{j,r_j} | \mathbf{y}, \mathbf{X}, \sigma_y^2, \tau, \phi_{j,r_j}, \mathbf{W}_{j,r_j} \sim$$
$$\mathrm{N}\left(\left(\frac{1}{\tau \phi_{j,r_j}} \mathbf{W}^{-1}_{j,r_j} + \frac{1}{\sigma_y^2} \mathbf{X}^{*\prime} \mathbf{X}^* \right)^{-1} \frac{\tilde{\mathbf{y}}' \mathbf{X}^*}{\sigma_y^2}, \left(\frac{1}{\tau \phi_{j,r_j}} \mathbf{W}^{-1}_{j,r_j} + \frac{1}{\sigma_y^2} \mathbf{X}^{*\prime} \mathbf{X}^* \right)^{-1} \right)$$

$\underline{\tau}$

$$p(\tau | a_\tau, b_\tau, \boldsymbol{\phi}, \mathbf{W}, \mathbf{B}) \propto \tau^{a_\tau - \frac{1}{2} \sum_{j=1}^D R_j p_j - 1} e^{-b_\tau \tau - \frac{1}{2\tau} \sum_{j=1}^D \sum_{r_j=1}^{R_j} \frac{1}{\phi_{j,r_j}} \boldsymbol{\beta}'_{j,r_j} \mathbf{W}^{-1}_{j,r_j} \boldsymbol{\beta}_{j,r_j}}$$
$$\tau | a_\tau, b_\tau, \boldsymbol{\phi}, \mathbf{W}, \mathbf{B} \sim \mathrm{GIG}\left(a_\tau - \frac{1}{2} \sum_{j=1}^D R_j p_j, 2b_\tau, \sum_{j=1}^D \sum_{r_j=1}^{R_j} \frac{1}{\phi_{j,r_j}} \boldsymbol{\beta}'_{j,r_j} \mathbf{W}^{-1}_{j,r_j} \boldsymbol{\beta}_{j,r_j} \right),$$

where GIG is shorthand for the Generalized Inverse Gaussian distribution.

$\underline{\pi_{j,r_j}}$

$$p(\pi_{j,r_j}|-) \propto \pi_{j,r_j}^{-p_j/2}(1-\pi_{j,r_j})^{\alpha-(R_j-r_j)p_j/2-1}\times \exp\left\{-\frac{1}{\tau}\left[\frac{1}{\pi_{j,r_j}}\left(\boldsymbol{\beta}'_{j,r_j}\mathbf{W}_{j,r_j}^{-1}\boldsymbol{\beta}_{j,r_j}\right) + \sum_{k=r_j+1}^{R_j}\frac{1}{\pi_{j,k}\prod_{\ell=r_j}^{k-1}(1-\pi_{j,\ell})}\left(\boldsymbol{\beta}'_{j,k}\mathbf{W}_{j,k}^{-1}\boldsymbol{\beta}_{j,k}\right)\right]\right\}$$

$\underline{\omega_{j,r_j,\ell}}$

$$\omega_{j,r_j,\ell}|\beta_{j,r_j,\ell},\tau,\phi_{j,r_j} \sim \text{Generalized Inverse Gaussian}\left(\frac{1}{2},\lambda_{j,r_j}^2,\frac{\beta_{j,r_j,\ell}^2}{\tau\phi_{j,r_j}}\right)$$

$\underline{\lambda_{j,r_j}}$

In order to find a posterior full conditional distribution in a closed form for λ_{j,r_j}, $\omega_{j,r_j,\ell}$ must be integrated out of the prior for $\beta_{j,r_j,\ell}$.

$$\int_0^\infty \frac{1}{\sqrt{2\pi\tau\phi_{j,r_j}\omega_{j,r_j,\ell}}}\exp\left\{-\frac{\beta_{j,r_j,\ell}^2}{2\tau\phi_{j,r_j}\omega_{j,r_j,\ell}}\right\}\frac{\lambda_{j,r_j}^2}{2}\exp\left\{-\frac{1}{2}\lambda_{j,r_j}^2\omega_{j,r_j,\ell}\right\}d\omega_{j,r_j,\ell} = \frac{\lambda_{j,r_j}}{2(\tau\phi_{j,r_j})^{-1/2}}\exp\left\{-\frac{\lambda_{j,r_j}|\beta_{j,r_j,\ell}|}{(\tau\phi_{j,r_j})^{-1/2}}\right\}$$

$$p(\lambda_{j,r_j}|\boldsymbol{\beta}_{j,r_j},\tau,\phi_{j,r_j}) \propto \lambda_{j,r_j}^{a_\lambda-1}\exp\left\{-b_\lambda\lambda_{j,r_j}\right\}\times\prod_{\ell=1}^{p_j}\lambda_{j,r_j}\exp\left\{-\frac{\lambda_{j,r_j}|\beta_{j,r_j,\ell}|}{(\tau\phi_{j,r_j})^{1/2}}\right\}$$

$$\lambda_{j,r_j}|\boldsymbol{\beta}_{j,r_j},\tau,\phi_{j,r_j} \sim \text{Gamma}\left(a_\lambda+p_j, b_\lambda+\frac{\sum|\beta_{j,r_j,\ell}|}{(\tau\phi_{j,r_j})^{1/2}}\right)$$

Bibliography

Armagan, A., Dunson, D. B., and Lee, J. (2013a). Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013b). Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):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*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bercovich, E. and Javitt, M. C. (2018). Medical imaging: From Roentgen to the digital revolution, and beyond. *Rambam Maimonides medical journal*, 9(4).

Bowman, F. D., Caffo, B., Bassett, S. S., and Kilts, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage*, 39(1):146–156.

Bro, R. (2006). Review on multiway analysis in chemistryâĂŤ2000–2005. *Critical reviews in analytical chemistry*, 36(3-4):279–293.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):627–641.

Brown, R. W., Cheng, Y.-C. N., Haacke, E. M., Thompson, M. R., and Venkatesan, R. (2014). *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons.

Bruno, M. A., Walker, E. A., and Abujudeh, H. H. (2015). Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I., Rousseau, J., et al. (2015a). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383.

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015b). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

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(4):2069–2101.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chumbley, J. R. and Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70.

Clyde, M., Desimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91(435):1197–1208.

Collins, D. L., Holmes, C. J., Peters, T. M., and Evans, A. C. (1995). Automatic 3-D model-based neuroanatomical segmentation. *Human brain mapping*, 3(3):190–208.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica*, 20(3):927–960.

Das, A., Sampson, A. L., Lainscsek, C., Muller, L., Lin, W., Doyle, J. C., Cash, S. S., Halgren, E., and Sejnowski, T. J. (2017). Interpretation of the precision matrix and its application in estimating sparse brain connectivity during sleep spindles from human electrocorticography recordings. *Neural computation*, 29(3):603–642.

Descombes, X., Kruggel, F., and Von Cramon, D. Y. (1998). Spatio-temporal fMRI analysis using Markov random fields. *IEEE transactions on medical imaging*, 17(6):1028–1039.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905.

Flandin, G. and Penny, W. D. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, 34(3):1108–1125.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., and Frackowiak, R. S. (1995). Spatial registration and normalization of images. *Human brain mapping*, 3(3):165–189.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., and Turner, R. (1998). Event-related fMRI: characterizing differential responses. *Neuroimage*, 7(1):30–40.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gerard, D. and Hoff, P. (2015). Adaptive higher-order spectral estimators. *arXiv preprint arXiv:1505.02114*.

Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562.

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., and Collins, D. L. (2006). Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 58–66. Springer.

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(79):1–31.

Hans, C. (2009). Bayesian LASSO regression. *Biometrika*, 96(4):835–845.

Huettel, S. A., Song, A. W., McCarthy, G., et al. (2004). *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., et al. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *Neuroimage*, 80:360–378.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik*, 31(1):253–258.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage*, 62(2):782–790.

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156.

Kalus, S., Sämann, P. G., and Fahrmeir, L. (2014). Classification of brain activation via spatial Bayesian variable selection in fMRI regression. *Advances in Data Analysis and Classification*, 8(1):63–83.

Kiers, H. A. and Mechelen, I. V. (2001). Three-way component analysis: Principles and illustrative application. *Psychological methods*, 6(1):84.

Kook, J. H., Guindani, M., Zhang, L., and Vannucci, M. (2017). NPBayes-fMRI: Non-parametric Bayesian general linear models for single-and multi-subject fMRI data. *Statistics in Biosciences*, pages 1–19.

Lazar, N. (2008). *The statistical analysis of functional MRI data*. Springer Science & Business Media.

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis (Online)*, 9(3):699.

Li, H. and Pati, D. (2017). Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107–119.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, pages 1–16.

Li, X., Xu, D., Zhou, H., and Li, L. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545.

Li, Y., Zhu, H., Shen, D., Lin, W., Gilmore, J. H., and Ibrahim, J. G. (2011). Multiscale adaptive regression models for neuroimaging data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):559–578.

Lindquist, M. A., Loh, J. M., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198.

Lindquist, M. A. and Mejia, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic medicine*, 77(2):114.

Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., and Poline, J.-B. (2003). Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Human Brain Mapping*, 19(1):1–17.

Martin, R., Mess, R., Walker, S. G., et al. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., et al. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322.

Mejia, A. F., Yue, Y., Bolin, D., Lindgren, F., and Lindquist, M. A. (2019). A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association*, pages 1–20.

Miller, L. and Milner, B. (1985). Cognitive risk-taking after frontal or temporal lobectomyâĂŤii. the synthesis of phonemic and semantic information. *Neuropsychologia*, 23(3):371–379.

Muschelli, J., Sweeney, E., Lindquist, M., and Crainiceanu, C. (2015). fslr: Connecting the FSL software with R. *The R Journal*, 7(1):163–175.

Nestor, S. M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J. L., Fogarty, J., Bartha, R., and Initiative, A. D. N. (2008). Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454.

Ng, B., Abugharbieh, R., Varoquaux, G., Poline, J. B., and Thirion, B. (2011). Connectivity-informed fMRI activation detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–292. Springer.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487.

Pangman, V. C., Sloan, J., and Guse, L. (2000). An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. *Applied Nursing Research*, 13(4):209–213.

Patel, R. S., Bowman, F. D., and Rilling, J. K. (2006a). A Bayesian approach to determining connectivity of the human brain. *Human brain mapping*, 27(3):267–276.

Patel, R. S., Bowman, F. D., and Rilling, J. K. (2006b). Determining hierarchical functional networks from auditory stimuli fMRI. *Human brain mapping*, 27(5):462–470.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538.

Polson, N. G. and Sun, L. (2019). Bayesian l 0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731.

Potts, R. B. (1952). Some generalized order-disorder transformations. In *Mathematical proceedings of the Cambridge Philosophical Society*, volume 48, pages 106–109. Cambridge University Press.

Quirós, A., Diez, R. M., and Gamerman, D. (2010). Bayesian spatiotemporal model of fMRI data. *NeuroImage*, 49(1):442–456.

Sanyal, N. and Ferreira, M. A. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *NeuroImage*, 63(3):1519–1531.

Schonberg, T., Fox, C. R., Mumford, J. A., Congdon, E., Trepel, C., and Poldrack, R. A. (2012). Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an fMRI investigation of the balloon analog risk task. *Frontiers in neuroscience*, 6.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Similä, T. and Tikka, J. (2007). Input selection and shrinkage in multiresponse linear regression. *Computational Statistics & Data Analysis*, 52(1):406–422.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102(478):417–431.

Smith, S. (2002a). Fast robust automated brain extraction.

Smith, S. M. (2002b). Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219.

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

Stefano, S., Quartagno, M., Tamburini, M., and Robinson, D. (2018). *orcutt: Estimate Procedure in Case of First Order Autocorrelation*.

Strittmatter, W. J. and Roses, A. D. (1996). Apolipoprotein E and Alzheimer's disease. *Annual review of neuroscience*, 19(1):53–77.

Sun, W. W. and Li, L. (2017). STORE: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944.

Sweeney, E., Crainiceanu, C. M., and Muschelli, J. I. (2014). *Introduction to Neurohacking in R*. Course attended Fall 2019.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Van Der Pas, S., Kleijn, B., Van Der Vaart, A., et al. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

Van der Vaart, A. W. and Van Zanten, H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675.

Van der Vaart, A. W. and Van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6):434–442.

Wang, H. et al. (2012). Bayesian graphical LASSO models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.

Wang, X., Nan, B., Zhu, J., and Koeppe, R. (2014). Regularized 3D functional regression for brain image data via Haar wavelets. *The Annals of Applied Statistics*, 8(2):1045.

Wang, Y., Kang, J., Kemmer, P. B., and Guo, Y. (2018). *DensParcorr: Dens-Based Method for Partial Correlation Estimation in Large Scale Brain Networks*. R package version 1.1.

Warnick, R., Guindani, M., Erhardt, E., Allen, E., Calhoun, V., and Vannucci, M. (2018). A Bayesian approach for estimating dynamic functional network connectivity in fMRI data. *Journal of the American Statistical Association*, 113(521):134–151.

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

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., and Rosseel, Y. (2011). neuRosim: An R package for generating fMRI data. *Journal of Statistical Software*, 44(10):1–18.

Welvaert, M. and Rosseel, Y. (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. *PloS one*, 8(11):e77089.

Whittemore, A. S. (2007). A Bayesian false discovery rate for multiple testing. *Journal of Applied Statistics*, 34(1):1–9.

Xu, L., Johnson, T. D., Nichols, T. E., and Nee, D. E. (2009). Modeling inter-subject variability in fMRI activation location: a Bayesian hierarchical spatial model. *Biometrics*, 65(4):1041–1051.

Yu, C.-H., Prado, R., Ombao, H., and Rowe, D. (2018). A Bayesian variable selection approach yields improved detection of brain activation from complex-valued fMRI. *Journal of the American Statistical Association*, pages 1–16.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.

Zhang, J., Li, X., Li, C., Lian, Z., Huang, X., Zhong, G., Zhu, D., Li, K., Jin, C., Hu, X., et al. (2014a). Inferring functional interaction and transition patterns via dynamic Bayesian variable partition models. *Human brain mapping*, 35(7):3314–3331.

Zhang, L., Guindani, M., and Vannucci, M. (2015). Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):21–41.

Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., and Vannucci, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *The Annals of Applied Statistics*, 10(2):638–666.

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014b). A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage*, 95:162–175.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.

Zhu, H., Fan, J., and Kong, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109(507):1084–1098.