Bayesian Regression for Skewed Tensor Response

Inkoo Lee

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

BAYESIAN REGRESSION FOR SKEWED TENSOR RESPONSE

By

INKOO LEE

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2021

Inkoo Lee defended this dissertation on March 26, 2021.
The members of the supervisory committee were:

Debajyoti Sinha

Professor Directing Dissertation

Sachin Shanbhag

University Representative

Qing Mai

Committee Member

Xin Zhang

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

Dedicated to my parents.

ACKNOWLEDGMENTS

I would like to convey my deepest gratitude to my advisor Dr. Debajyoti Sinha for his advice and encouragement. It is a big fortune to have the excellent mentor to illuminate great ideas in Bayesian and ensure my work when I lost a direction. I am thankful to him for the tremendous faith and patience he invested in me. I am also grateful to Dr. Qing Mai who always had an open door and give intuitive feedback when I asked myriad of questions about the paper. I would also like to thank Dr. Xin Zhang for his insights into the tensor regression problem and for providing me valuable comments and unreserved help. It has been an honor to work with them and this dissertation would not be the same without their diligence. I would also appreciate my committee member Dr. Sachin Shanbhag for his suggestions and willingness to set aside time for my defenses.

Outside of my committee, I would thank to Dr. Dipankar Bandyopadhyay for providing GAAD data which motivates my research and great advice as well. Also, I am grateful to Dr. Jonathan Bradley for providing the initial inspiration for me to start the journey of research.

Dr. George Rust has provided the great opportunity to participate the collaborative work to tackle racial disparity problem with medical school professors Dr. Henry Carretta, Dr. Yi Luo, and medical student Gabrielle LeBlanc. It has been my privilege to work with you all. I would express a deep appreciation to Dr. Rust for supporting me under the Florida DOH Bankhead-Coley Cancer Research Grant. I would also acknowledge that this research was supported by Hobbs Foundation of Research and Dissertation Research Grant from the Graduate School.

In addition, I would especially like to thank my peers Chenran Wang, Zishen Xu, Qing Han, Yinpu Li and Jingze Liu. They brought me great memories and happy moments in the last five years.

Lastly, I would like to thank my parents and Seunghee Choi for their patience and support, without whose support I would not have had the opportunity to succeed. I would like to acknowledge with appreciation to my brother for his faith in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A substantial amount of work exists for tensor regression analysis in a variety of clinical settings, including neuroimaging, genomics and dental medicine. Our motivation for this paper is from a study of periodontal disease (PD) with a three-dimensional tensor response: multiple *biomarkers* measured at pre-specified *tooth sites* within each *tooth*, for each subject. A careful investigation would reveal considerable skewness in the responses, in addition to response missingness. To mitigate the shortcomings of existing multivariate regression tools (that ignore the inherent tensor structure) and tensor normal based methods (that ignore response skewness), we propose a new Bayesian tensor response regression method that facilitates interpretation of covariate effects on *both* marginal and joint distributions of the tensor response, accommodating missing responses under a closure property. Furthermore, we present a prudent evaluation of the overall covariate effects, as well as identifying their possible variations on only a sparse subset of the tensor components. Our method promises MCMC tools that are readily implementable. We illustrate substantial advantages of our estimation proposal over existing methods via simulation studies, and application to the PD dataset.

We propose a general class of skewed elliptical distributions for tensor responses to ensure that any linear combinations of tensor variables still follow tensor elliptical distributions. Additionally, the marginal density has the same form as the conditional density of skewed tensor elliptical distribution. Our class of skewed elliptical distributions has useful properties including the following: the class is closed under marginalization and includes skewed tensor normal and skewed tensor-t distributions as special cases. Exploiting tensor form provides multiple benefits; 1) we can use maximum information of tensor structure that cannot be employed by multivariate methods, 2) tensor covariance structure captures the dependence of each direction of tensor. Practical applications of this new class are provided via Bayesian tensor response regression analysis with two types of prior for tensor regression coefficients: tensor normal (TN) prior, and tensor spike-and-slab lasso (TSSL) prior. We illustrate the practical advantages of TSSL prior to detecting fast decaying teeth types and sites in a periodontal disease study.

CHAPTER 1

INTRODUCTION

1.1 Motivations

To date, multidimensional array (tensor) which can be found in biomedical studies and associated applications is a common trait of modern complex data. However, relatively few works have been conducted on statistical tools to analyze tensor data; much of today's literature focuses on one-dimensional (vector) data analysis in which researchers collapse multidimensional data into vectors for simplicity.

Another example of simplification in tensor data analysis is the common assumption that the data follows a Tensor Normal (TN) distribution when there is a large enough number of observations. This is a naïve assumption especially unsuitable for highly skewed tensor data, which commonly occurs in real life even with large sample sizes. These methods lead to the use of limited information resulting in either biased estimation or inaccurate prediction. Therefore, our central goal is to develop a new class of tensor distributions with theoretical framework in order to bridge the gap between the regression analysis of tensor responses and highly skewed data with possible missingness.

The motivating example of this dissertation is the periodontal disease (PD) data of Gullah-speaking African-American diabetics, (henceforth, GAAD study) [19]. The PD data consists of 28 teeth 6 surfaces and 2 bio-markers per one patient; two highly skewed biomarkers: the periodontal pocket depth (PPD) and clinical attachment level (CAL), were measured (in mm) by dental hygienist at available tooth-sites, within each tooth of a patient. PPD measures current PD status, while CAL is the gold standard for diagnosis and monitoring PD measuring past progression and disease history of PD [41].

In our studies, we introduce two types of skewed tensor distributions; skewed tensor normal (STN) (in chapter 2) and skewed tensor-t (STT) distributions (in chapter 3). Furthermore, we develop the general theoretical framework for skewed tensor elliptical class (STEL). Note that STN and STT distributions are special cases of STEL. We also discuss the useful properties of STEL

and their special cases. We demonstrate superior performances of our models compared to existing methods via simulation studies and GAAD data analyses.

1.2 Literature Review

Statistical methodologies for tensor data analyses have been widely developed in past decades. Particularly, many tensor regression models have been introduced with several scenarios such as 1) tensor covariate 2) tensor response.

We first review tensor covariate regression models. Within frequentist literature, [57] proposed generalized linear model (GLM) framework with tensor covariate exploiting the PARAFAC decomposition. [56] introduced tensor partial least squares algorithms achieving sufficient dimension reduction with tensor envelope and then construct population interpretation. [34] employed Tucker decomposition [52] and introduced low rank approximation for tensor coefficient. Along this direction, Bayesian approach using low-rank structure of PARAFAC decomposition for tensor coefficient was proposed by [22]. Furthermore, [42] developed soft PARAFAC decomposition by varying specific entries in the row contributions and it provides extra flexibility for (hard) PARAFAC.

In contrast, the literature of regression of tensor response on a vector of covariates is comparatively scant. In frequentist paradigm, [33] proposed envelope methods, and [50] introduced a low-rank decomposition and element-wise sparsity. In the same vein, [23, 49] developed Bayesian tensor regression models deploying PARAFAC decomposition to estimate low-ranked tensor regression coefficients. Also, [21] developed a generalized Bayesian regression framework with a symmetric tensor response and scalar predictors. They embedded low-rankness and group sparsity on the symmetric tensor coefficients.

In a multivariate setting, there exists a vast body of literature [1, 2, 3, 5, 7, 10, 48] to avoid possible risks of ignoring non-Gaussian structure when practitioners analyze the highly skewed data. Multivariate framework contorts structure of tensor data and one may only use limited information. So, [24] provided general class of skewed matrix-variate elliptical distributions. However, the regression literature dealing with highly skewed tensor response and its theoretical framework is still very much in its infancy. Therefore, this dissertation tackle this problem.

1.3 Tensor Notations and Operations

We briefly present the necessary notations and operations, mostly adapted from [32]. We use lower-case letters c for scalar, lower-case bold letters $\mathbf{v}$ for vector, upper bold letters $\mathbf{M}$ for matrix, and calligraphic capital letters $\mathcal{T}$ for tensor. A multidimensional array $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_K}$ is defined as a *tensor* of order K (number of *modes* of $\mathcal{Y}$). So, vectors and matrices are first and second order tensors, respectively. A *mode-m fiber* $\mathcal{Y}_{i_1 \cdots i_{m-1} : i_{m+1} \cdots i_K}$ is defined by fixing every index of the tensor, except the indexes of m-th mode. The $\text{vec}(\mathcal{Y})$ operator stacks all $\prod_{k=1}^{K} d_k$ elements of $\mathcal{Y}$ within this column vector, so that the element $y_{i_1 \ldots i_K}$ of the tensor $\mathcal{Y}$ is the j-th entry of $\text{vec}(\mathcal{Y})$, where $j = 1 + \sum_{m=1}^{K} (i_m - 1) \prod_{m'=1}^{m-1} d_{m'}$. The *mode-m matricization*, $\mathbf{Y}_{(m)}$, is a $d_m \times (\prod_{j \neq m} d_j)$ dimensional matrix, such that the $(i_1, \ldots, i_K)$-th element of tensor $\mathcal{Y}$ is the (i_k, j)-th element of the matrix $\mathbf{Y}_{(m)}$ for $j = 1 + \sum_{m \neq m'} (i_{m'} - 1) \prod_{m'' < m', m'' \neq m} d_{m''}$. The *m-mode product* $\mathcal{Y} \times_m \mathbf{S}$ of tensor $\mathcal{Y}$ with a matrix $\mathbf{S} \in \mathbb{R}^{W \times d_m}$ is a K-way tensor with dimension $(d_1 \times \cdots \times d_{m-1} \times W \times d_{m+1} \times \cdots \times d_K)$, where the $(i_1, \cdots, i_{m-1}, w, i_{m+1}, \cdots, i_K)$-th element of $\mathcal{Y} \times_m \mathbf{S}$ is $\sum_{i_m=1}^{d_m} y_{i_1 i_2 \cdots i_K} s_{w i_m}$. When $\mathbf{s}$ is a d_m-dimensional vector, $\mathcal{Y} \bar{\times}_m \mathbf{s}$ reduces to a $(K-1)$-way tensor with size $\mathbb{R}^{d_1 \times \cdots \times d_{m-1} \times d_{m+1} \times \cdots \times d_K}$. The Kronecker product of two matrices is denoted, $\mathbf{A} \otimes \mathbf{B}$. The *Tucker decomposition* of a tensor $\mathcal{G}$, denoted by $\mathcal{G} = [\![\mathcal{H}; \mathbf{V}_1, \ldots, \mathbf{V}_K]\!]$, is defined as $\mathcal{G} = \mathcal{H} \times_1 \mathbf{V}_1 \times_2 \cdots \times_K \mathbf{V}_K$, where $\mathcal{H} \in \mathbb{R}^{u_1 \times \cdots \times u_K}$ is called core tensor, and $\mathbf{V}_k \in \mathbb{R}^{d_k \times u_k}$ for $k = 1, \ldots, K$ are called factor matrices. We also utilize the fact that $\text{vec}([\![\mathcal{H}; \mathbf{V}_1, \ldots, \mathbf{V}_K]\!]) = (\mathbf{V}_K \otimes \cdots \otimes \mathbf{V}_1)\text{vec}(\mathcal{H})$. The inner product of two tensors $\mathcal{Y}$ and $\mathcal{Y}$ of same size is defined as $\langle \mathcal{Y}, \mathcal{Y} \rangle = \text{vec}(\mathcal{Y})^\intercal \text{vec}(\mathcal{Y})$.

A tensor random variable $\mathcal{C} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ has a standard tensor normal distribution when all the entries of $\mathcal{C}$ are independent standard normal. So, $\mathcal{Y} = \mathcal{M} + [\![\mathcal{C}; \mathbf{\Sigma}_1^{1/2}, \ldots, \mathbf{\Sigma}_K^{1/2}]\!]$, has a tensor normal distribution denoted by $\mathcal{Y} \sim TN(\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K)$, where the positive definite matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}_k^{1/2} \mathbf{\Sigma}_k^{1/2}$ models the dependence structure on the kth-mode. In this case, $\text{vec}(\mathcal{Y}) = \text{vec}(\mathcal{M}) + \mathbf{\Sigma}^{1/2} \text{vec}(\mathcal{C})$ has the multivariate normal distribution, $\text{vec}(\mathcal{Y}) \sim N_{d^*}(\text{vec}(\mathcal{M}), \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1$, and $d^* = \prod_{k=1}^{K} d_k$.

CHAPTER 2

BAYESIAN REGRESSION ANALYSIS OF SKEWED TENSOR RESPONSES

2.1 Introduction

In recent times, multidimensional array (tensor) responses are becoming commonplace in biomedical studies, and other applications [33, 22]. For example, in the cross-sectional periodontal disease (PD) study of Gullah-speaking African-American diabetics, henceforth, GAAD study [19], the tensor response are two biomarkers: the periodontal pocket depth (PPD) and clinical attachment level (CAL), measured at pre-specified tooth-sites, within each tooth of a study participant. The PPD and CAL measure respectively, the (current) PD status, and (past) progression and disease history of PD [41], and hence should be modeled jointly while quantifying covariate effects. However, existing analyses of PD responses mostly either ignore or contort the original tensor response structure of these studies. For example, [7] and [4] averaged the observed responses over teeth and sites to reduce them to bivariate summary responses leading to imminent loss of information. Very recently, [55] reduced the site-level PPD and CAL responses as two columns in their matrix variate response modeling. For a proper understanding of covariate effects, we ought to utilize the original tensor structure of the biomarker responses.

The analysis of GAAD study offer other challenges. For example, a common practice for regression analysis of tensor response is to use a tensor normal distribution [6], which does not accommodate skewness of the distribution of the tensor response. However, both PPD and CAL responses high degree of skewness (see panel (a-b) of Figure 2.1). Even after separately fitting linear mixed-effects (LME) model with Gaussian assumptions of the random effects and random errors [4] to these responses using `R nlme4`, the self-calibrated Q-Q plots of the residuals of PPD and CAL from LME models (panels e and f of Figure 2.1) reveal clear evidence of asymmetry, i.e., departures from Gaussian error assumptions. However, those non-symmetric features are not so prominent for the empirical Bayes' estimates of the random effects (panels c and d). Given the difficulty to

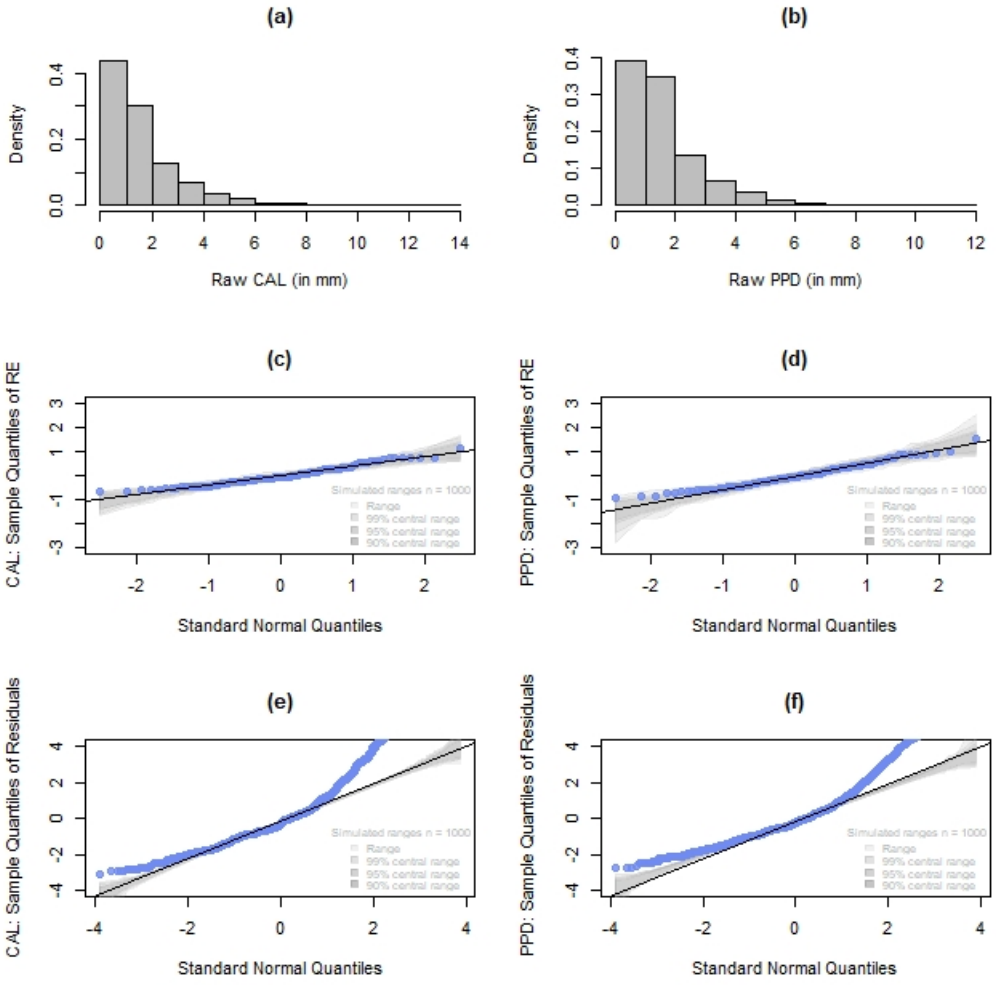

Figure 2.1: GAAD Data: Plots of the histogram of the raw CAL and PPD responses (panels a and b), and the corresponding self-calibrated Q-Q plots of the empirical Bayes' estimates of the random effect (panels c and d) and of the error residuals (panels e and f) for visual testing of the null hypothesis of Gaussianity at various significance levels, obtained after fitting linear mixed models to the PPD and CAL responses, separately, controlling for all covariates, using `R` packages `lme4` and `qqtest`.

decide an appropriate transformation to normality for a tensor response data, the challenge is to use a suitable model and associated analysis for highly skewed tensor responses.

An extensive body of literature [3, 48, 2, 5, 10] exists to emphasize the perils of analyzing highly skewed responses while ignoring their non-Gaussian structure, especially in a multivariate regression setting. Furthermore, the responses exhibit missingness (about 32.4% missing teeth), which has been hypothesized to be mostly due to past incidence of PD, and analyzed via shared random-effects models [46] under non-ignorable, missing-not-at-random (MNAR) assumptions. However, assumptions of non-ignorability require a model for the missingness process, and may lead to issues in identifiability [36].

Many of the existing tensor regression models, primarily aimed at neuroimaging analysis [57, 56, 22], deal with the regression of clinical responses on a tensor image. In contrast, the literature of regression of tensor response on a vector of covariates is relatively sparse. Recently, [33] introduced envelope methods [13], [50] proposed a low-rank decomposition and [23] proposed a sparse Bayesian method to deal with high dimensional tensor responses. For our motivating GAAD study the relevant analysis goal is to evaluate and interpret the overall effects of scalar covariates on tensor response as well as to identify possibly sparse subset of tensor components that experience very different covariate effects compared to the rest while accommodating the highly skewed nature of these tensor responses. To bridge the gap between the regression analysis of tensor responses and highly skewed data with possible missingness, we develop a new class of tensor distribution that we call *skewed tensor normal*. Our model accommodates the skewnesses of tensor responses via *tensor skewing shocks*, and the corresponding Bayesian treatment under missing-at-random assumptions is termed as *Bayesian skewed tensor normal* (BSTN) regression. A major advantage of our model is that it accommodates two separate sets of parameters to handle skewness and variability, with Bayesian learning enabled by specifying two independent prior choices on them, based on separate marginal prior knowledge about skewness and variability of these tensor responses. Accordingly, our method allows the interpretation of the skewness level and covariate effects for each component of the tensor response.

This article provides several contributions. First, our model is closed under marginalization to facilitate the interpretation of the covariate effects on any subset of tensor responses. Second, unlike some existing methods for skewed multivariate responses [3], our method can evaluate sep-

arate effects of covariate and skewness on all marginal densities of the tensor responses. Third, missing responses are not unusual in PD datasets, and our computationally convenient method for handling missing tensor responses (under missing-at-random assumptions) is the first such attempt in the area of tensor regression. Finally, we utilize closed form inverses and determinants of covariance matrices to ease computation. The association structure and sparse modeling of regression coefficient employed here are different from the Gaussian noise assumptions and the PARAFAC decomposition based tensor coefficient modeling used widely in the tensor regression literature [33, 23, 49]. Furthermore, our Markov Chain Monte Carlo (MCMC) tools are implementable in `R`.

The remainder of the article proceeds as follows. After an introduction to tensor notations and operations, we introduce our new skewed tensor response regression model in Section 2.2. In Section 2.3, we outline the Bayesian method including the likelihood accommodating missing responses, hierarchical prior specification, and the posterior computation. To evaluate the finite sample performances of the BSTN method and comparisons to existing alternatives, we present simulation studies with synthetic data generated from a variety of scenarios in Section 2.4. In Section 2.5, we illustrate our BSTN model via application to the GAAD data. Finally, we conclude with a discussion and future research directions in Section 2.6. Technical results and proofs are relegated within Appendices.

2.2 Statistical Model

2.2.1 Skewed Tensor Response Model

For a K-th order tensor response $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ with a vector of covariates $\mathbf{x}_i \in \mathbb{R}^p$, we consider a tensor response regression model

$$\mathcal{Y}_i = \mathcal{B} \bar{\times}_{(K+1)} \mathbf{x}_i + \mathcal{E}_i, \quad \text{for} \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathcal{B} \in \mathbb{R}^{d_1 \times \cdots \times d_K \times p}$ is an $(K+1)$th order unknown tensor of regression coefficients, $\bar{\times}_{(K+1)}$ is the $(K+1)$-mode vector product, and the error $\mathcal{E}_i \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a Kth order tensor. We model the skewness in the distribution of $\mathcal{Y}$ via

$$\mathcal{E}_i = |\mathcal{Z}_{2i}| \times_K \mathbf{\Lambda} + \mathcal{Z}_{1i}, \tag{2.2}$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{d_K}) \in \mathbb{R}^{d_K \times d_K}$ and skewness parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_K})$. We use $|\mathbf{M}|$ to denote the operation where each element of $|\mathbf{M}|$ is the absolute value of the corresponding element of $\mathbf{M}$. In (2.2), the tensor skewing shock $\mathcal{Z}_{2i} \sim TN(\mathbf{0}; \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}_{\boldsymbol{\sigma}}^2)$ is assumed to independent of $\mathcal{Z}_{1i} \sim TN(\mathbf{0}; \mathbf{R}_1, \ldots, \mathbf{R}_{K-1}, \mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}})$, where $\mathbf{R}_k > 0$ is a correlation matrix for $k = 1, \ldots, K$, $\mathbf{D}_{\boldsymbol{\sigma}} = \text{diag}(\sigma_1, \ldots, \sigma_{d_K})$ with $\sigma_j > 0$ for $j = 1, \ldots, d_K$. When $\boldsymbol{\lambda} = \mathbf{0}$, then $\mathcal{E}_i$ of (2.2) equals to $\mathcal{Z}_{1i} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ having a tensor normal distribution. We denote the tensor distribution of $\mathcal{E}_i$ in (2.2) as the skewed tensor normal distribution $\mathcal{E}_i \sim STN(\mathbf{0}; \mathbf{R}_1, \ldots, \mathbf{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$ and the corresponding distribution of $\mathcal{Y}_i$ in (2.1) as $\mathcal{Y}_i \sim STN(\mathcal{B}\bar{\times}_{(K+1)}\mathbf{x}_i; \mathbf{R}_1, \ldots, \mathbf{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$. The $\mathbf{R}_1, \ldots, \mathbf{R}_K$ are correlation matrices (instead of usual covariance matrices used for tensor normal densities) to ensure the identifiability of the model parameters in (2.2) for the typical studies we consider for data analysis.

2.2.2 Important Model Properties

We now present some important properties of STN model of (2.1)-(2.2). These properties are useful to identify the parameters governing the marginal skewness and scales of each tensor component, and correlation between a pair of components. The covariance matrix $\text{cov}\{\text{vec}(\mathcal{Z}_{1i})\} = \mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}} \otimes \mathbf{R}_{K-1} \otimes \cdots \otimes \mathbf{R}_1$ has separable covariance structure [26, 27, 33]. As shown later, this separable covariance structure reduces computational cost of our Bayesian method. The model in (2.1) implies

$$\text{vec}(\mathcal{Y}_i) = \mathbf{B}_{(K+1)}^{\intercal}\mathbf{x}_i + (\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 \ldots d_{K-1}})\text{vec}(|\mathcal{Z}_{2i}|) + \text{vec}(\mathcal{Z}_{1i}), \quad \text{for} \quad i = 1, \ldots, n, \tag{2.3}$$

where $\text{vec}(\mathcal{Y}_i) \in \mathbb{R}^{d^*}$ for $d^* = \prod_{k=1}^{K} d_k$, $\mathbf{B}_{(K+1)} \in \mathbb{R}^{p \times d^*}$ is the mode-$(K+1)$ matricization (see notations earlier) of the tensor of regression coefficients $\mathcal{B}$ in (2.1). Each column of $\mathbf{B}_{(K+1)}$ in (2.3) is a vector of coefficients describing the linear relationship between the individual elements of $\mathcal{Y}_i$ and the covariate $\mathbf{x}_i$. The benefit of the model in (2.1)-(2.2) is that the covariate effects, scale and skewness parameters have separate marginal interpretations. We will use a special case of the STN model of (2.1)-(2.2) with $K = 3$, $d_1 = 28$ teeth, $d_2 = 6$ sites and $d_3 = 2$ biomarkers for analyzing the GAAD study.

For tensor distribution in (2.2), the skewness is isotropic in the first $K - 1$ modes, while the skewness $\boldsymbol{\lambda} \in \mathbb{R}^{d_K}$ is specified *only* on the K-th mode. So, every d_k-dimensional mode-K fiber

$\mathcal{E}_{i_1\cdots i_{K-1}:}$ of error $\mathcal{E}$ has the common distribution $\mathcal{E}_{i_1\cdots i_{K-1}:} \sim MSN(\mathbf{0}, \mathbf{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$ which is the Multivariate Skew Normal distribution of [7] with representation

$$\mathcal{E}_{i_1\cdots i_{K-1}:} = \boldsymbol{\Lambda}|\mathbf{z}_2| + \mathbf{z}_1 \text{ for independent } \mathbf{z}_1 \sim N_{d_K}(\mathbf{0}, \mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}}),\ \mathbf{z}_2 \sim N_{d_K}(\mathbf{0}, \mathbf{D}^2_{\boldsymbol{\sigma}}). \tag{2.4}$$

The skewness of tensor error $\mathcal{E}_i$ in (2.2) is modeled via latent tensor "skewing shocks" $|\mathcal{Z}_{2i}|$ (similar to skewing shocks $|\mathbf{z}_{2i}|$ for fiber $\mathcal{E}_{i_1\cdots i_{K-1}:}$ in (2.4)), and the skewness parameters $\boldsymbol{\lambda}$. Also, when λ_j is positive (or negative), the corresponding marginal density of $y_{i_1,\ldots,i_{K-1},j}$ of tensor response $\mathcal{Y}$ is skewed to the right (left). For example, the 3-way (that is, $K = 3$) tensor response $\mathcal{Y}$ from a participant in the GAAD study with $d_1 = 28$ teeth, $d_2 = 6$ sites for each tooth, and $d_3 = 2$ biomarkers (PPD and CAL), the model in (2.2) assumes that each element of $(28 \times 6 \times 2)$ tensor response $\mathcal{Y}$ has independent skewing shock, but, the level of skewness is same for all elements of the tensor response corresponding to the same biomarker (λ_1 for all PPD responses and λ_2 for all CAL responses). As a consequence, the error distributions corresponding to all elements of the response for the same biomarker are equal, with mean $\mathrm{E}(e_{i_1,\ldots,i_{K-1},j}) = \lambda_j \mathrm{E}(|z_{2,i_1,\ldots,i_{K-1}j}|) + \mathrm{E}(z_{1,i_1,\ldots,i_{K-1}j}) = \lambda_j\sigma_j\sqrt{2/\pi}$ and variance $\mathrm{Var}(\mathcal{E}_{i_1\ldots i_{K-1},j}) = \lambda_j^2\mathrm{Var}(|z_{2,i_1,\ldots,i_{K-1}j}|) + \mathrm{Var}(z_{1,i_1,\ldots,i_{K-1}j}) = \lambda_j^2\sigma_j^2(1-2/\pi) + \sigma_j^2 = \sigma_j^2[1+\lambda_j^2\{1-(2/\pi)\}]$ for $j = 1, \ldots, d_K$. These results also demonstrate that the skewness parameter λ_j and the scale parameter σ_j for model in (2.2) have very different roles for the distribution of the j-th biomarker of any site of any tooth.

The correlation between any pair of responses within subject i is

$$\mathrm{corr}(e_{i,i_1\ldots i_K}, e_{i,i'_1\ldots i'_K}) \ = \ \left\{\prod_{k=1}^{K}\rho_{i_k i'_k}\right\}\left[\left\{1+\lambda_{i_K}^2\left(1-\frac{2}{\pi}\right)\right\}\left\{1+\lambda_{i'_K}^2\left(1-\frac{2}{\pi}\right)\right\}\right]^{-\frac{1}{2}} . \tag{2.5}$$

It is worth noting that the correlation in (2.5) is a function of correlation matrices $\mathbf{R}_1, \ldots, \mathbf{R}_K$ (via $\rho_{i_k i'_k}$ for $k = 1, \cdots, K$) and skewness parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_K})$, but it is free of the scale parameters $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{d_K})$. The Pearson's first skewness coefficient $\{\mathrm{E}(e_{i,i_1,\ldots,i_K}) - \mathrm{mode}(e_{i,i_1,\ldots,i_K})\}/\sqrt{\mathrm{Var}(e_{i,i_1,\ldots,i_K})}$ [44] of $\mathcal{E}_{i,i_1\ldots i_K}$ in (2.2) is equal to $\lambda_{i_K}\sqrt{2/\pi}/[1+\lambda_{i_K}^2\{1-(2/\pi)\}]^{1/2}$ because the $\mathrm{mode}(\mathcal{E}_{i,i_1,\ldots,i_K}) = 0$. As desired, this skewness coefficient of our response distribution is only a function of λ_{i_K}, and does not depend on σ_{i_K} and $\mathbf{R}_k$.

2.3 Bayesian Inference

For the tensor response $\mathcal{Y}_i$ from $i = 1, \ldots, n$ independent participants (subjects), the likelihood function $L(\Theta|\mathcal{Y})$ of the parameters $\Theta = (\mathcal{B}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \boldsymbol{\sigma})$ of model (2.1)-(2.2) is proportional to

$$\begin{aligned}\prod_{i=1}^{n} f(\mathcal{Y}_i|\Theta) \propto \prod_{i=1}^{n} \int f_1(\mathcal{Y}_i - \mathcal{B}\bar{\times}_{(K+1)}\mathbf{x}_i - |\mathcal{Z}_{2i}| \times_K \boldsymbol{\Lambda}\big|\mathbf{R}_1, \ldots, \mathbf{R}_{K-1}, \mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}}) \\ \times f_2(\mathcal{Z}_{2i}|\mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}^2_{\boldsymbol{\sigma}})d\mathcal{Z}_{2i},\end{aligned} \tag{2.6}$$

where f_1 represents the density of $TN(\mathbf{0}; \mathbf{R}_1, \ldots, \mathbf{R}_{K-1}, \mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}})$ of $\mathcal{Z}_{1i}$, and f_2 represents $TN(\mathbf{0}; \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}^2_{\boldsymbol{\sigma}})$ density of latent tensor $\mathcal{Z}_{2i}$ in (2.2).

2.3.1 Incorporating Missing Responses

Several subjects in the GAAD study have missing responses in some components of the $d_1 \times d_2 \times d_3$ tensor $\mathcal{Y}_i$ [46], with missingness a commonplace in PD monitoring studies. To easily accommodate these missing responses within the likelihood, we use the desirable property that our model is closed under marginalization over covariate effects and skewness levels. We define the $d_1 \times d_2 \times d_3$ missing data indicator tensor $\mathcal{U}_i$, with entries $u_{ii_1i_2i_3} = 0$ if $y_{ii_1i_2i_3}$ is observed, $u_{ii_1i_2i_3} = 1$ if $y_{ii_1i_2i_3}$ is missing. By abuse of notation, we denote $\mathcal{Y}_{i,c} = \{\mathcal{Y}_{\text{i,obs}}, \mathcal{Y}_{\text{i,mis}}\}$, where $\mathcal{Y}_{\text{i,obs}}$ is observed tensor response and $\mathcal{Y}_{\text{i,mis}}$ represents the missing part. Under either missing at random (MAR), or missing completely at random (MCAR) assumptions, we can ignore the missing mechanism (distribution of $\mathcal{U}_i$), and the likelihood contribution from $\mathcal{Y}_{\text{i,obs}}$ in (2.6) is proportional to $f(\mathcal{Y}_{i,\text{obs}}|\Theta) = \int f(\mathcal{Y}_{\text{i, obs}}, \mathcal{Y}_{\text{i,mis}}|\Theta)d\mathcal{Y}_{\text{i,mis}}$.

We now present Theorem 1 that ensures that we can easily obtain the likelihood contribution $f(\mathcal{Y}_{\text{i,obs}}|\Theta)$ of any observed tensor response $\mathcal{Y}_{\text{i,obs}}$ without integrating out the $\mathcal{Y}_{\text{i,mis}}$ numerically.

Theorem 1. *Suppose the K-th tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$ of dimension $\mathcal{I} := d_1 \times \cdots \times d_K$ has the distribution $\mathcal{Y} \sim STN(\mathcal{B}\bar{\times}_{(K+1)}\boldsymbol{x}, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$ for some $\boldsymbol{x} \in \mathbb{R}^p$ and $(K+1)$-th order tensor $\mathcal{B}$ of dimension $\mathcal{I}_2 := d_1 \times \cdots \times d_K \times p$, and $\mathcal{Y}^{(m)}$ is a K-th order tensor of dimension $\mathcal{I}^* := d_1 \times \cdots d_{m-1} \times d^*_m \times d_{m+1} \cdots \times d_K$ such that $\mathcal{Y}^{(m)}_{i_1 \cdots, i_K} = \mathcal{Y}_{i_1 \cdots, i_K}$ when $1 \le i_1 \le d_1, \cdots, 1 \le i_m \le d^*_m, \cdots, 1 \le i_K \le d_K$, then this sub-tensor of $\mathcal{Y}$ has the distribution $\mathcal{Y}^{(m)} \sim STN(\mathcal{B}^{(m)}\bar{\times}_{(K+1)}\boldsymbol{x}; \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{m-1}, \boldsymbol{R}^*_m, \mathbf{R}_{m+1}, \ldots, \boldsymbol{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$, where $\mathbf{R}_m$ is partitioned as $\begin{pmatrix} \boldsymbol{R}^*_m & \boldsymbol{R}_{m1} \\ \boldsymbol{R}_{m1} & \boldsymbol{R}_{m2} \end{pmatrix}$ for $\boldsymbol{R}^*_m \in \mathbb{R}^{d^*_m \times d^*_m}$, and $\mathcal{B}^{(m)}$ is the corresponding $(K+1)$-th order tensor of*

dimension $\mathcal{I}_3 := d_1 \times \cdots \times d_{m-1} \times d_m^* \times d_{m+1} \times \cdots \times d_K \times p$ *such that* $\mathcal{B}^{(m)}_{i_1\ldots,i_kj} = \mathcal{B}_{i_1\ldots,i_kj}$ *for* $1 \leq i_m \leq d_m^*$ *and* $1 \leq i_k \leq d_k$ *when* $k \neq m$. *For the boundary case of* $m = K$ *above,* $\mathcal{Y}^{(K)} \sim STN(\mathcal{B}^{(K)} \bar{\times}_{(K+1)} \boldsymbol{x}; \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{K-1}, \boldsymbol{R}_K^*; \boldsymbol{\sigma}^*, \boldsymbol{\lambda}^*)$, *where* $\boldsymbol{\sigma}^*$ *and* $\boldsymbol{\lambda}^*$ *are corresponding vectors of first* d_K^* *elements of* $\boldsymbol{\sigma}, \boldsymbol{\lambda}$.

The proof of Theorem 1 for the case $m = 1$ follows from the constructive definition of the skew-tensor-normal distribution in (2.1)-(2.2) and fact that if a K-way tensor $\mathcal{Z}$ follows a tensor normal distribution with parameters $(\mathcal{M}, \mathbf{R}_1, \ldots, \mathbf{R}_K)$, then a corresponding sub-tensor $\mathcal{Z}^{(1)}$ also follows a tensor normal distribution with parameters $(\mathcal{M}^{(1)}, \mathbf{R}_1^*, \mathbf{R}_2, \ldots, \mathbf{R}_K))$ with $\mathcal{M}^{(1)}$ and $\mathbf{R}_1^*$ being defined according in relation to $\mathcal{M}$ and $\mathbf{R}_1$. Without of loss of generality, this proof can now be straightforwardly extended to the cases when $1 < m \leq K$. We skip the details of the proof for the sake of brevity. Theorem 1 shows that the class of skewed tensor distribution in (2.1)-(2.2) is closed under marginalization, when we partition the tensor response $\mathcal{Y}_i$ over a specific mode. Particularly, it is relevant for GAAD study where any missing tooth (mode-1) of a patient has all its sites and biomarkers missing. For the GAAD study, the observed response distribution $\mathcal{Y}_{i,\text{obs}} \sim STN(\mathcal{B}^{(1)} \bar{\times}_{(K+1)} \mathbf{x}_i; \mathbf{R}_{11}, \mathbf{R}_2, \mathbf{R}_3, \boldsymbol{\sigma}, \boldsymbol{\lambda})$ (from Theorem 1) used for likelihood contribution in (2.6) via the corresponding partition of $(\mathcal{B}, \mathbf{R}_1)$.

2.3.2 Joint Posterior

The joint posterior density is given by

$$p(\Theta|\mathcal{Y}_0) \propto L(\Theta|\mathcal{Y})\pi(\mathcal{B})\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\sigma})\pi(\boldsymbol{\rho}) \ , \tag{2.7}$$

where, the observed data likelihood $L(\Theta|\mathcal{Y}_0)$ is the likelihood in (2.6) except the likelihood contribution $f(\mathcal{Y}_i|\Theta)$ should be based on observed data response distribution $\mathcal{Y}_{iobs} \sim STN(\mathcal{B}^{(1)} \bar{\times}_{(K+1)} \mathbf{x}_i; \mathbf{R}_{11}, \mathbf{R}_2, \mathbf{R}_3, \boldsymbol{\sigma}, \boldsymbol{\lambda})$ using Theorem 1. By abuse of notation. $\pi(\cdot)$ represents independent marginal prior densities of $\boldsymbol{\lambda}, \boldsymbol{\sigma}$, and $\boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is the vector of unknown parameters of the parametric correlation matrices $(\mathbf{R}_1, \ldots, \mathbf{R}_K)$. In the context of the motivating GAAD study with a three-way tensor response $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and a vector of covariates (including intercept) $\mathbf{x}_i \in \mathbb{R}^p$, we now specify details of the priors on different components of Θ and associated MCMC steps for the posterior in (2.7) . The property that the Pearson's first skewness coefficient of the marginal distribution of each component $\mathcal{Y}_{i_1 i_2 i_3}$ depends only on λ_{i3} allows us to specify the independent priors for the

skewness level $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ and scale parameter $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$, according to the marginal opinions about the skewness and scale of each of two biomarkers in GAAD study.

2.3.3 Prior Specification

With our 3-way tensor response $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, and a vector of covariates (including intercept) $\mathbf{x}_i \in \mathbb{R}^p$, we now specify the priors on the components of the parameter vector Θ in our model (2.1)-(2.2) as follows.

(i) For tensor coefficient $\mathcal{B}$, we first use a tensor normal $\pi(\mathcal{B})$, say, $TN(\mathbf{0}; \mathbf{C}_1, \mathbf{C}_2,$
$\mathbf{C}_3, \mathbf{C}_4)$ with prior mean zero tensor with known covariance matrices $\mathbf{C}_1, \ldots, \mathbf{C}_4$. This prior modeling allows every element $\beta_{i_1 i_2 i_3, j}$ to be potentially different from each other.

(ii) For the skewness parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ in GAAD study, we use independent mean zero normal with pre-specified variance as the common prior for for each biomarker.

(iii) For the scale parameters $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ corresponding to PPD and CAL in the GAAD study, we use common independent inverse-gamma distribution denoted by $IG(g_1, g_2)$, with the pre-specified shape $g_1 > 0$ and scale $g_2 > 0$.

(iv) The parametric correlation matrices $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$ in GAAD study are assumed to be equicorrelation matrices with unknown off-diagonal element ρ_k for $k = 1, 2, 3$. However, our method can allow other parametric correlation matrices for $\mathbf{R}_k$. We use independent Uniform$(-1, 1)$ as the common prior for ρ_k.

We would like to emphasize that these above parameterizations of $\mathbf{R}_k$ and prior specifications can be arguably extended to other parametric forms of correlation matrices (example, unrestricted) and other appropriate prior distributions. We implement MCMC sampling following the conditional posterior distributions in (2.8). We sample $\mathcal{B}$, λ_{i_K}, $\sigma^2_{i_K}$, ρ_k and $z_{i,i_1 \ldots i_K}$ using a Gibbs sampler for $k = 1, \ldots, K$ and $i_K = 1, \ldots, d_K$. In absence of a standard closed form for the kernel of the posterior distribution of $\boldsymbol{\rho}$, we update $\boldsymbol{\rho}$ using Metropolis-Hastings algorithm [11]. The next subsection outlines our stepwise MCMC implementation.

2.3.4 Posterior Computation

With $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, we proceed as follows:

$$\text{Step 1}: \quad p(\text{vec}(\mathcal{B})|-) \sim N\big(\text{A}^{-1}\big\{\text{vec}(S_{xy}) - \text{vec}(S_{xw})\big\}, \text{A}^{-1}\big)$$

where $A = \big(S_{xx} \otimes \mathbf{\Sigma}^{\star -1} + \frac{1}{\mathrm{c}^3}\mathbf{I}_{d_1 d_2 d_3 p}\big)$, $S_{xy} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i^{\intercal} \mathbf{\Sigma}^{\star -1}$,

$S_{xw} = \sum_{i=1}^{n} \mathbf{x}_i \mathrm{vec}(|\mathcal{Z}_{2i}|)^{\intercal}(\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 d_2})\mathbf{\Sigma}^{\star -1}$, and $S_{xx} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\intercal}$.

Step 2 : $p(\lambda_{i_3}|-) \sim N\Big(\frac{G_\delta}{H_{\delta,\delta}}, \frac{1}{H_{\delta,\delta}}\Big)$,

where $H_{\delta,\delta} = \sum_{i=1}^{n} \big[(|\mathbf{Z}_{2i(3)}|^{\intercal} \otimes \mathbf{I}_{d_3})^{\intercal} \mathbf{\Sigma}^{\star -1} (|\mathbf{Z}_{2i(3)}|^{\intercal} \otimes \mathbf{I}_{d_3}) + \mathbf{I}_{d_3^2}\big]_{{i_3}^2, {i_3}^2}$,

and $G_\delta = \sum_{i=1}^{n} \big\{\big[(\mathbf{y}_i^{\intercal} - \mathbf{x}_i \mathbf{B}_{(4)}) \mathbf{\Sigma}^{\star -1} (|\mathbf{Z}_{2i(3)}|^{\intercal} \otimes \mathbf{I}_{d_3}) + 1\big]_{{i_3}^2}$

$- \sum_{{i_3}^2 \neq {i_3}'^2} \big[(|\mathbf{Z}_{2i(3)}|^{\intercal} \otimes \mathbf{I}_{d_3})^{\intercal} \mathbf{\Sigma}^{\star -1} (|\mathbf{Z}_{2i(3)}|^{\intercal} \otimes \mathbf{I}_{d_3}) + \mathbf{I}_{d_3^2}\big]_{{i_3}^2, {i_3}'^2} \lambda_{i_3'}$, for $i_3 = 1,\ 2$.

Step 3 : $p\Big(\frac{1}{\sigma_j^2}\Big| - \Big) \sim Ga\Big(nTS + g_1, \nu\Big)$, where j = 1, 2, Ga denotes Gamma distribution.

$$\nu = \frac{1}{2} \sum_{i=1}^{n} \Bigg\{ \mathbf{S}^{\intercal} \Bigg[\begin{pmatrix} \frac{1}{(1-\rho_3)(1+\rho_3)} & -\frac{\sigma_1 \rho_3}{\sigma_2 (1-\rho_3)(1+\rho_3)} \\ -\frac{\sigma_1 \rho_3}{\sigma_2 (1-\rho_3)(1+\rho_3)} & \frac{\sigma_1^2}{\sigma_2^2 (1-\rho_3)(1+\rho_3)} \end{pmatrix} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1} \Bigg] \mathbf{S}$$

$$+ \mathrm{vec}(|\mathcal{Z}_{2i}|)^{\intercal} \mathrm{diag}(1, \sigma_1^2/\sigma_2^2, \ldots, 1, \sigma_1^2/\sigma_2^2) \mathrm{vec}(|\mathcal{Z}_{2i}|) \Bigg\} + g_2,$$

and $\mathbf{S} = \mathrm{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^{\intercal} \mathbf{x}_i - (\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 d_2}) \mathrm{vec}(|\mathcal{Z}_{2i}|)$.

Step 4 : We use Metropolis-Hastings algorithm to sample $\boldsymbol{\rho}$. Assume that $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$.

The target density of ρ_1 is following

$$p(\rho_1|-) \underset{\rho_1}{\propto} \big\{1/((1-\rho_1)^{d_1 - 1}(1 + (d_1 - 1)\rho_1))\big\}^{d_2 d_3}\big]^{n/2} \times$$

$$\exp\Bigg[-\frac{1}{2} \sum_{i=1}^{n} \mathbf{S}^{\intercal} \Bigg\{ \mathbf{D}_{\boldsymbol{\sigma}}^{-1} \mathbf{R}_3^{-1} \mathbf{D}_{\boldsymbol{\sigma}}^{-1} \otimes \mathbf{R}_2^{-1} \otimes \frac{1}{1-\rho_1} \Big(\mathbf{I}_{d_1} - \frac{\rho_1}{1 + (d_1 - 1)\rho_1} \mathbf{1}_{d_1 \times d_1} \Big) \Bigg\} \mathbf{S} \Bigg] \mathbf{I}_{(0,1)}(\rho_1).$$

The proposal density is beta distribution with shape parameter h_1 and scale parameter h_2, where $h_1 \geq h_2 \geq 2$. The support of beta distribution is between 0 and 1 which is same as support of ρ_1. Similarly, beta distribution is proposal density for ρ_2 and ρ_3.

Step 5 : $p\big(|z_{2i,i_1 i_2 i_3}| - \big) \sim TrN\Big(\frac{E_\eta}{D_{\eta,\eta}}, \frac{1}{D_{\eta,\eta}}\Big) I(|z_{2i,i_1 i_2 i_3}| > 0)$,

where TrN stands for (univariate) truncated normal distribution with support $(0, \infty)$,

$D_{\eta,\eta} = \big\{(\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 d_2}) \mathbf{\Sigma}^{\star -1} (\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 d_2}) + (\mathbf{I}_{d_1 d_2} \otimes \mathbf{D}_{\boldsymbol{\sigma}}^2)^{-1}\big\}_{i_1 i_2 i_3, i_1 i_2 i_3}$, and

$E_\eta = \{(\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 \ldots d_2}) \mathbf{\Sigma}^{\star -1} (\mathbf{y}_i - \mathbf{B}_{(4)}^{\intercal} \mathbf{x}_i)\}_{i, i_1 i_2 i_3}$

$- \sum_{i_1 i_2 i_3 \neq i_1' i_2' i_3'} \big\{(\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 d_2}) \mathbf{\Sigma}^{\star -1} (\mathbf{\Lambda} \otimes \mathbf{I}_{d_1 d_2}) + (\mathbf{I}_{d_1 d_2} \otimes \mathbf{D}_{\boldsymbol{\sigma}}^2)^{-1}\big\}_{i_1 i_2 i_3, i_1' i_2' i_3'} \mathrm{z}_{2i, i_1' i_2' i_3'}$.

(2.8)

Details on the derivations of the full conditional distributions, and the associated MCMC steps (combining the Metropolis-Hastings and Gibbs sampling) are provided in the Appendices. We use `R` to implement the MCMC sampling algorithm.

2.3.5 Sparse Tensor Prior and Posteriors

The tensor normal prior on $\mathcal{B}$ in (2.1) puts no constraint on $p\prod_{k=1}^{K} d_k$ elements of $\mathcal{B}$. In practice, we need substantial reduction of dimension of the regression coefficients for the sake of practical interpretation and theory. One existing method for such a dimension reduction is via the PARAFAC decomposition of $\mathcal{B}$ along with corresponding sparsity inducing priors [49, 23]. In spite of such PARAFAC decomposition based priors being very useful in certain applications, the class of sparsity inducing priors for $\mathcal{B}$ in practice should depend on the study and its analysis goals. For studies such as GAAD, there is no guarantee of the existence of the low-rank decomposition of $\mathcal{B}$, and there is no available prior opinion about the order of such low-rank reduction needed for PARAFC decomposition based analysis. Also, the PARAFAC decomposition of $\mathcal{B}$ is not unique, and the identifiability of such a decomposition is not guaranteed even with constraints of known norms for all except one of the margin vectors ([32]). To accommodate a practical interpretation of the prior on $\mathcal{B}$ for the analysis of the studies such as GAAD, we present a novel sparsity inducing prior on $\mathcal{B}$ to model overall regression effects of each covariate on two biomarkers, as well as to identify only small subset of the teeth and sites that experience covariate effects different from the rest. For example, in the GAAD study, previous research [46] has shown that molar sites are usually more prone to periodontal decay than rest and especially, say, incisors. Even within molars, certain sites may decay faster over age than the others molar sites. To achieve this, we use

$$\beta_{i_1 i_2 i_3 j} = \eta_{i_3 j} + \gamma_{i_1 i_2 i_3 j} \text{ with prior } \eta_{i_3 j} \sim N(0, \tau^2) \text{ for } \tau > 0 \tag{2.9}$$

and introduce sparsity vis spike-and-slab priors [29] as

$$\gamma_{i_1 i_2 i_3 j} \overset{iid}{\sim} (1 - \omega_{i_1 i_2 i_3 j})\delta_0(\gamma_{i_1 i_2 i_3 j}) + \omega_{i_1 i_2 i_3 j} N(0, \nu^2) \text{ for } \nu > 0, \tag{2.10}$$

where, $\delta_0(\cdot)$ is the indicator function at 0 and $\omega_{i_1 i_2 i_3 j}$ has the hyperprior $\omega_{i_3 j} \overset{iid}{\sim} Bernoulli(\psi)$, with $\psi \sim Beta(a, b)$. For the GAAD study, we choose $a = b = 0.1$, representing a vague hyperprior for ψ. If the sparse prior (2.10) eventually selects only M non-zero $\gamma_{i_1 i_2 i_3 j}$, then we have only $(M + d_3 p)$

non-zero parameters, instead of $pd_1d_2d_3$ elements of unrestricted $\mathcal{B}$ with tensor Normal prior. We later demonstrate its practical advantages particularly for analysis of GAAD study.

We denote $\boldsymbol{\gamma}_k = \text{vec}(\boldsymbol{\Gamma}_k)$ and $\boldsymbol{\eta}_3 = \text{vec}(\boldsymbol{H}_3)$, where $\boldsymbol{\Gamma}_k = (\boldsymbol{\gamma}_{k,1}, \ldots, \boldsymbol{\gamma}_{k,p})$ for $k = 1, 2$ and $\boldsymbol{H}_3 = (\boldsymbol{\eta}_{3,1}, \ldots, \boldsymbol{\eta}_{3,p})$, for $\boldsymbol{\eta}_{3,j} = (\eta_{1j}, \ldots, \eta_{d_3 j})^\intercal$ for $j = 1, \ldots, p$. We now present the conditional posterior distributions for $\boldsymbol{\gamma_k}, \boldsymbol{\eta}_3, \omega_{i_1 i_2 i_3 j}, \psi$ given rest in (2.11).

$$
\begin{aligned}
&\boldsymbol{\gamma}_k|- \sim (1-\boldsymbol{\omega}_k)\delta_0(\boldsymbol{\gamma}_k) + \boldsymbol{\omega}_k N(\mathrm{A}_k^{-1}\{\text{vec}(S_{kxy}) - \text{vec}(S_{kxw})\}, \mathrm{A}_k^{-1}), \text{ for } k = 1, 2, \text{ and} \\
&\mathrm{A}_k = 1/\sigma^2\{S_{xx} \otimes \mathbf{R}_k^{-1}\} + 1/\nu_k^2 \mathbf{I}_{d_k p}, \; S_{kxy} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_{ki}^\intercal \mathbf{R}_k^{-1}, \; S_{kxw} = \sum_{i=1}^{n} \mathbf{x}_i |\mathbf{z}_{2,ki}|^\intercal \lambda \mathbf{R}_k^{-1} \; ; \\
&\boldsymbol{\eta}_3|- \sim N(\mathrm{A}_3^{-1}\{\text{vec}(S_{3xy}) - \text{vec}(S_{3xw})\}, A_3^{-1}), \text{ where } \mathrm{A}_3 = S_{xx} \otimes \mathbf{D}_{\boldsymbol{\sigma}}^{-1}\mathbf{R}_3^{-1}\mathbf{D}_{\boldsymbol{\sigma}}^{-1} + 1/\tau_k^2 \mathbf{I}_{d_3 p}, \\
&S_{3xy} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_{3i}^\intercal \mathbf{D}_{\boldsymbol{\sigma}}^{-1}\mathbf{R}_3^{-1}\mathbf{D}_{\boldsymbol{\sigma}}^{-1}, \; S_{3xw} = \sum_{i=1}^{n} \mathbf{x}_i |\mathbf{z}_{2,3i}|^\intercal \lambda \mathbf{D}_{\boldsymbol{\sigma}}^{-1}\mathbf{R}_3^{-1}\mathbf{D}_{\boldsymbol{\sigma}}^{-1}, \; S_{xx} = \sum_{i=1}^{n} \mathbf{x}_i^\intercal \mathbf{x}_i \; ; \\
&\omega_{i_k j}|- \sim Bernoulli(\zeta_{i_k j}) \text{ for } \zeta_{i_k j} = \frac{\psi \nu^{-d_k p} \det(\mathrm{A}_k^{-1})^{1/2} \exp\{\frac{1}{2\sigma^2}\text{vec}(\mathrm{S}_k)^\intercal \mathrm{A}_k^{-1} \text{vec}(\mathrm{S}_k)\}}{[1 - \psi + \psi \nu^{-d_k p} \det(\mathrm{A}_k^{-1})^{1/2} \exp\{\frac{1}{2\sigma^2}\text{vec}(\mathrm{S}_k)^\intercal \mathrm{A}_k^{-1} \text{vec}(\mathrm{S}_k)\}]}, \\
&\text{vec}(S_k) = \text{vec}(S_{kxy}) - \text{vec}(S_{kxw}) \text{ for } k = 1, 2; \\
&\psi|- \sim Beta\left(a + \sum_{i_k=1}^{d_k}\sum_{j=1}^{p} \omega_{i_k j}, b + \sum_{i_k=1}^{d_k}\sum_{j=1}^{p}(1 - \omega_{i_k j})\right).
\end{aligned}
\tag{2.11}
$$

2.4 Simulation Study

The goal of our simulation study is to compare the accuracy of our proposed method with other competing methods when the true $\mathcal{B}$ has neither a low-rank decomposition nor a sparsity structure. The tensor responses $\mathcal{Y}_i$ for $i = 1, \ldots, n$ are simulated from two different tensor models: our Skew Tensor Normal (STN) model of (2.2), and the Tensor Normal (TN) model (a special case of STN) with no skewness. We consider two different Bayesian tensor regression methods: the Bayesian Tensor Normal (BTN) without accommodating skewness, and the Bayesian Skewed Tensor Normal (BSTN), accommodating tensor skewing shocks of (2.2). We also consider comparing existing methods, such as the (a) ordinary least squares (OLS), that fits one response component at a time disregarding the tensor structure, and (b) the envelope-based tensor regression model (ENV) of [33], with estimated dimensions of envelope (tuning parameter). We do not compare

with PARAFAC/CP method of [23] due to the lack of reasonable prior opinion on the rank of the PARAFAC decomposition to have a fair comparison with the rest of the methods.

We simulate three-way $(10 \times 3 \times 2)$ tensor response data from the model: $\mathcal{Y}_i = \mathcal{B}\bar{\times}_4\mathbf{x}_i + |\mathcal{Z}_{2i}| \times_3 \mathbf{\Lambda} + \mathcal{Z}_{1i}$, with order-4 tensor coefficients $\mathcal{B}$. We generate half of the true $\mathcal{B}$ coefficients from $N(1,1)$, and the rest from $N(-1,1)$. This ensures that approximately half of tensor coefficients are positive, and the rest are negative. We simulate the corresponding covariate vector $\mathbf{x}_i = (1, x_{1i})^\intercal$ as $x_{1i} \sim Ber(0.5)$. The goal is to verify whether the BSTN based estimates almost perform as good as BTN based estimates when the true simulation model is Tensor Normal (scenario i), and they perform better than all competing methods when the true simulation model is skew tensor normal (scenario ii). We are also interested in evaluating the performance of BSTN method regarding estimating the skewness parameters. The true skewness parameters for the Skew Tensor Normal simulation model (scenario ii) used $\boldsymbol{\lambda} = (1, 2)$, while the Tensor Normal (scenario i) simulation model has no skewness. For scenario (ii), tensor skewing shocks $|\mathcal{Z}_{2i}|$ were simulated based on equicorrelation matrices $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$ with common off-diagonals $\rho_1 = \rho_2 = \rho_3 = 0.6$, and $\sigma_1^2 = \sigma_2^2 = 1$. Note, the BTN, OLS and ENV methods do not estimate the skewness parameters. We used the improper flat prior $TN(\mathbf{0}; 10\mathbf{I}_{10}, 10\mathbf{I}_3, 10\mathbf{I}_2, 10\mathbf{I}_2)$ on $\mathcal{B}$ for BSTN and BTN based analyses of simulated data. It is important to note that the prior for $\mathcal{B}$ used in the simulation study do not offer any particular advantage to Bayesian methods. However, in practice, the prior for $\mathcal{B}$ should be chosen based on available prior opinions to benefit the inference in terms of the better precision. For example, for the Bayesian analysis of GAAD study, the sparsity inducing prior of $\mathcal{B}$ described in (2.9)) offer practical advantages.

We compare the tensor estimates $\hat{\mathcal{B}}$ from the competing methods based on the average MSE (Mean Square Errors) given by $E||\mathcal{B} - \hat{\mathcal{B}}||_F^2/120$ of $\hat{\mathcal{B}}$, wher $||\mathcal{T}||_F$ is the Frobenius norm of the tensor $\mathcal{T}$ and the expectation is with respect to the true sampling distribution of $\hat{\mathcal{B}}$ under the simulation model. Similarly, we evaluate the Bayes estimate $\hat{\lambda_k}$ from BSTN based on MSE: $E[(\lambda_k - \hat{\lambda_k})^2]$. Table 2.1 presents the approximated average MSE of $\hat{\mathcal{B}}$ from all methods, and approximate MSE of the estimated skewness parameters (only for BSTN method) for three different sample-sizes under each of two simulation models. The approximations of MSE are based on the replication size of 30. The results show a better performance of the BTN than other methods when the true model is tensor normal, while BSTN has superior performance compared to competitors when the true

model is skew tensor normal. However, performance of BSTN is very comparable to BTN even when the true model is TN (the difference in MSE of two methods decreases fast as the sample size grows). BTN and OLS show similar performances as the sample size increases, while BTN performs substantially better than OLS and ENV when the sample size is quite small ($n = 20$). Skewness levels are well-estimated by BSTN method under both scenarios, and are close to the true values as the sample size grows. Therefore, BSTN is a safer method for analyzing three-way tensor responses especially when the true model may have substantial skewness, and shows superior performance compared to others when the responses actually follow STN distribution.

Table 2.1: Simulation study results: The approximate MSE (sampling errors) of the estimates of $\mathcal{B}$, λ_1 and λ_2 from BTN, BSTN and other competing methods (OLS and ENV) for analysis of 3-way tensor response data simulated from Tensor Normal and Skewed Tensor Normal models, across different sample sizes choices. The lowest MSE for each simulation scenario is highlighted in boldface.

Simulation		BTN	BSTN			OLS	ENV
models	n	$\mathcal{B}$	$\mathcal{B}$	λ_1	λ_2	$\mathcal{B}$	$\mathcal{B}$
Tensor	20	**0.2069** (0.041)	0.2290 (0.048)	0.151 (0.14)	0.142 (0.13)	0.2172 (0.046)	0.2400 (0.111)
Normal	50	**0.0859** (0.012)	0.0906 (0.012)	0.097 (0.11)	0.090 (0.11)	0.0873 (0.013)	0.1060 (0.053)
	100	**0.0412** (0.007)	0.0428 (0.007)	0.073 (0.11)	0.062 (0.08)	0.0417 (0.008)	0.0605 (0.018)
Skewed	20	0.5293 (0.083)	**0.4925** (0.079)	0.054 (0.04)	0.047 (0.03)	0.5335 (0.083)	0.5742 (0.152)
Tensor	50	0.2078 (0.035)	**0.1979** (0.034)	0.047 (0.03)	0.045 (0.03)	0.2145 (0.036)	0.2341 (0.063)
Normal	100	0.0974 (0.012)	**0.0912** (0.011)	0.040 (0.02)	0.039 (0.02)	0.0982 (0.012)	0.1088 (0.034)

2.5 Application: GAAD Dataset

The GAAD study conducted by the Center for Oral Health Research (COHR) at the Medical University of South Carolina (MUSC) primarily aimed at exploring the relationship between PD and diabetes level in Type-2 diabetic Gullah-speaking African-Americans ($\geq$13 years old), residing in coastal South Carolina. To measure PD status/progression, dental hygienists recorded two biomarkers, the PPD and CAL [14], measured in mm for each of the 6 sites per tooth (disto-buccal, mid-buccal, mesio-buccal, disto-lingual, mid-lingual and mesio-lingual), for a maximum of 28 teeth per participants (excluding the 4 third molars). Note, for a missing tooth, both biomarkers from all its corresponding sites are missing.

Our central goal is to assess cross-sectional association of the 5 subject-specific covariates, Age (in years), Gender (1=Female, 0=Male), BMI (Body Mass Index, a measure of obesity with

obese when BMI $\geq$ 30), HbA1c [12] level (1 = uncontrolled, 0 = controlled), and Smoker (1 if current or previous smoker, 0 = never smoker), on PPD and CAL. For our analysis, we use $n = 288$ subjects with complete covariate information, and with at least one tooth available. So, for GAAD study participant i, the regression function for each element of the tensor response is $Y_{i_1,i_2,i_3} = \sum_{j=1}^{p} \beta_{i_1,i_2,i_3,j} x_{ij} + \mathcal{E}_{i_1 i_2 i_3}$ from (2.1). We consider three competing methods: the BSTN method with a skewed tensor normal density for $\mathcal{E}_i$ in (2.1), the BTN method with tensor normal density for $\mathcal{E}_i$, and the OLS (classical/frequentist) method. Missing responses were accommodated as MAR (described earlier). For Bayesian analysis using MCMC, we assess the convergence of 2 independent chains using trace plots, auto-correlation plots, and the Gelman-Rubin [20] diagnostics.

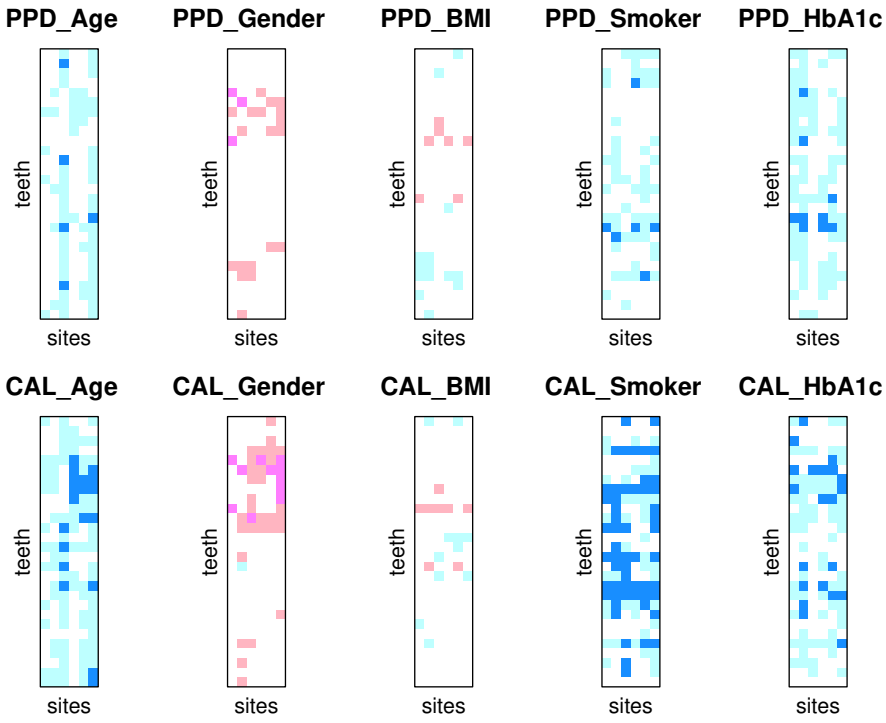

Figure 2.2: Fitting the BSTN Model with Tensor Normal prior for $\mathcal{B}$ to the GAAD data: Plotted are the D statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong and moderate evidences of the positive effects, strong and moderate evidences of negative effects and evidence to no effect are shaded respectively in dark and light blues, dark and light pinks and in white.

2.5.1 Using Non-Sparse $\mathcal{B}$

When the tensor regression effects $\mathcal{B}$ is not assumed to have any sparsity structure, we present the results of our analysis using the statistic $D = d_1/|d_2|$ for each regression effects $\mathcal{B}_{i_1 i_2 i_3 j}$, where d_1 is distance of the Bayesian estimate (posterior median) from null value (0 here) and d_2 is the distance of the Bayes estimate from the nearest endpoint of the 95% CI of the parameter. Here, $D > 1.5$ ($D < -1.5$) implies strong posterior evidence of positive (negative) effect of covariate x_j on the biomarker i_3 at site i_2 of tooth i_1. Similarly, the range $-0.8 < D < 0.8$ suggests moderate evidence for null hypothesis, and $0.8 \leq D \leq 1.5$ represents moderate positive effect, and $-1.5 \leq D \leq -0.8$ denotes moderate negative effect of x_j on $\mathcal{Y}_{i_1 i_2 i_3}$. Figure 2.2 presents the heatmaps of these D-statistics obtained from the BSTN model fit to the two biomarkers (two rows of rectangular color plates) for the various tooth-site (6 columns of small colored boxes) and tooth (28 rows of boxes within each plate) combinations. The plot illustrates the strength of evidence and the direction of association of each covariate on the PPD and CAL measures, for all 28×6 teeth-sites combinations. Even though Figure 2.2 demonstrates positive (in this case adverse) effects of Age, HbA1c and Smoking, and negative (beneficial) effects of gender (females are better) on the two biomarkers for most of the (tooth, site) combinations, the vastly checkerboard pattern makes it difficult to effectively summarize, and even comprehend the overall effects of these covariates, possibly due to the high-dimensionality of $\mathcal{B}$. Similar patterns were also observed in the D-statistics heatmaps from the fit of the BTN model with the tensor normal prior on $\mathcal{B}$, the OLS, and the ENV models, as displayed in Figures A.2–A.4, respectively, in the Appendices. Note that the D-statistics plots for the OLS and ENV method uses 1000 bootstrap samples for variance estimation.

For each of 5 covariates, Table A.3 (Appendices) present the average widths of the interval estimates of the (covariate) association measures for all 28×6 teeth-sites combinations, obtained from the BTN, BSTN (with TN prior on $\mathcal{B}$), and the OLS. Overall, the BSTN method produces more precise parameter estimates, compared to BTN and the OLS methods. Note, Bayesian methods provide interval estimates and posterior uncertainty of the regression coefficients directly, whereas, for the OLS, we employ resampling tools. Compared to both BTN and BSTN, the OLS method produces wider 95% interval estimates. It is interesting to note that even BSTN with $\mathbf{R}_3 = \mathbf{I}$ has smaller average width of interval estimates, compared to the BTN. Hence, incorporating skewness

is more crucial for estimation than allowing correlation structure in the third mode of the tensor responses (the two bio-markers in the GAAD study).

2.5.2 Using Sparse Priors on $\mathcal{B}$

For a more clinically informative yet parsimonious assessment of covariate effects at a further granular (tooth/site) level, we further conduct the Bayesian analysis by using the overall effect $\eta_{i_3 j}$ of covariate effect x_j on PPD and CAL ($i_3 = 1, 2$) and a sparsity inducing prior on the (site, tooth) specific effects $\gamma_{i_1 i_2 i_3 j}$, specified in subsection 2.3.5. Table 2.2 presents the posterior summaries (posterior median, standard deviations (SD) and the 95% credible interval (CI)) of the overall effects $\eta_{i_3 j}$, for the 5 covariates and intercept terms on PPD and CAL, from fitting the BSTN method with sparsity prior for $\mathcal{B}$. Age, Smoker and HbA1c have strong posterior evidences of positive (adverse) effects on PPD and CAL, however, the effects of BMI, though positive, do not have substantial data evidence (95% CI includes 0). The binary covariate Gender has substantial negative effects (beneficial effect for women). The estimates of λ_1 and λ_2 reveal significant skewness, with the skewness in CAL to be larger in magnitude than the PPD. Compared to the bivariate regression model using full-mouth average data in [7], the corresponding CIs for the skewness parameters and overall regression effects are narrower. The corresponding estimates from the BTN method (with sparsity prior) and the OLS are available in Table A.4 (Appendices). We observe that the 95% intervals for most parameters from these competing models are wider, compared to corresponding estimates from the BSTN methods with the sparsity prior.

The heatmap of D-statistics in Figure 2.3 obtained from fitting the BSTN method with sparsity prior display a clear pattern of the posterior evidence of covariate associations on PPD and CAL for all 28×6 individual tooth-site combinations, compared to the checkerboard pattern (Figure 2.2) from the corresponding fit without sparsity priors. We observe strong posterior evidence of the positive association of HbA1c on both biomarkers for almost all (tooth, site) combinations, with a more prominent association with CAL. Smoking and Age also exhibit similar strong evidence of positive associations on both PPD and CAL for most (tooth, site) combinations, with more evidence on CAL. Gender exhibits strong negative association for both PPD and CAL, implying PD status of females better than males. On the contrary, Figure 2.4 presenting the corresponding D-statistics heatmap from the BTN fit with sparsity prior fails to present a parsimonious summary

Table 2.2: Fitting the BSTN Model with sparse tensor prior for $\mathcal{B}$ to the GAAD data. Values in table are the posterior summaries of the overall covariate associations, and the skewness parameters, corresponding to the PPD (upper row), and CAL (lower row).

PPD	Median	SD	CI
Intercept	0.8287	0.1489	(0.5631, 1.0801)
Age	0.0081	0.0022	(0.0036, 0.0138)
Gender	-0.1104	0.0621	(-0.2106, -0.0462)
BMI	0.0124	0.0551	(-0.0969, 0.1128)
Smoker	0.1813	0.0588	(0.0786, 0.2997)
HbA1c	0.2048	0.0649	(0.0836, 0.3199)
Skewness	1.2004	0.0137	(1.1734, 1.2226)
CAL	Median	SD	CI
Intercept	-0.0289	0.1455	(-0.4400, 0.3531)
Age	0.0159	0.0024	(0.0105, 0.0222)
Gender	-0.1821	0.0684	(-0.3331, -0.0284)
BMI	0.0243	0.0609	(-0.1190, 0.1706)
Smoker	0.3364	0.0619	(0.1922, 0.4802)
HbA1c	0.2649	0.0697	(0.1141, 0.4076)
Skewness	1.5557	0.0101	(1.5347, 1.5751)

of the covariate associations, with often conflicting results, such as, negligible to no association of age with PPD.

Table 2.3: Fitting the BSTN model with sparsity to the GAAD data. Values presented are the percentages of the posterior medians (Bayes point estimates) of $\gamma_{i_1 i_2 i_3 j} \neq 0$, for 6 tooth-sites of interest (combined across tooth-types)

PPD	mesio-buccal	disto-buccal	mid-buccal	mesio-lingual	disto-lingual	mid-lingual
Age	100%	100%	100%	100%	100%	100%
Gender	71.43%	64.29%	57.14%	57.14%	60.71%	71.43%
BMI	75%	71.43%	60.71%	71.43%	57.14%	71.43%
Smoker	78.57%	85.71%	71.43%	75%	78.57%	53.57%
HbA1c	75%	71.43%	67.86%	64.29%	67.86%	60.71%
CAL	mesio-buccal	disto-buccal	mid-buccal	mesio-lingual	disto-lingual	mid-lingual
Age	100%	100%	100%	100%	100%	100%
Gender	64.29%	64.29%	71.43%	67.86%	75%	89.29%
BMI	57.14%	67.86%	78.57%	64.26%	64.26%	85.71%
Smoker	67.86%	71.43%	92.86%	64.29%	78.57%	75%
HbA1c	64.28%	78.57%	75%	60.71%	60.71%	75%

Furthermore, to detect and quantify varying covariate effects on the tensor responses (PPD and CAL) between tooth-sites (combined across tooth-types; see Table 2.3) and between tooth-types (combined across tooth-sites; see Table 2.4), we obtain the posterior summaries (specifically, the posterior median and 95% CI) of $\gamma_{i_1 i_2 i_3 j}$, the difference in the overall and the (tooth i_1, site i_2) specific association, corresponding to the j-th covariate on biomarker $i_3 = 1, 2$. Tables 2.3 and 2.4 present the percentages of the non-zero posterior median of $\gamma_{i_1 i_2 i_3 j}$, for the tooth-sites and tooth-

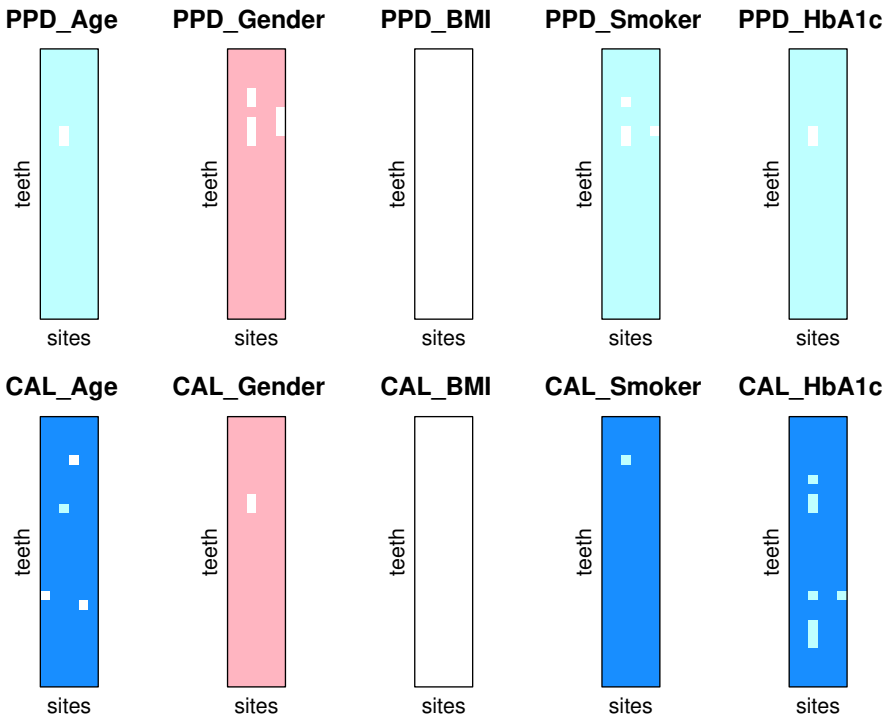

Figure 2.3: Fitting the BSTN Model with sparsity prior on $\mathcal{B}$ to the GAAD data. Plotted are the D statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

Table 2.4: Fitting the BSTN model with sparsity to the GAAD data. Values presented are the percentages of the posterior medians of $\gamma_{i_1 i_2 i_3 j} \neq 0$ for 4 tooth-types (combined across tooth-sites).

PPD	Molars	Pre-molars	Canines	Incisors
Age	100%	100%	100%	100%
Gender	81.25%	58.33%	45.83%	58.33%
BMI	83.33%	68.75%	58.33%	56.25%
Smoker	100%	70.83%	54.17%	60.42%
HbA1c	91.67%	70.83%	54.17%	50%
CAL	**Molars**	**Pre-molars**	Canines	Incisors
Age	100%	100%	100%	100%
Gender	81.25%	75%	62.5%	58.33%
BMI	89.58%	66.67%	62.5%	54.16%
Smoker	100%	81.25%	66.67%	50%
HbA1c	91.67%	75%	66.67%	47.92%

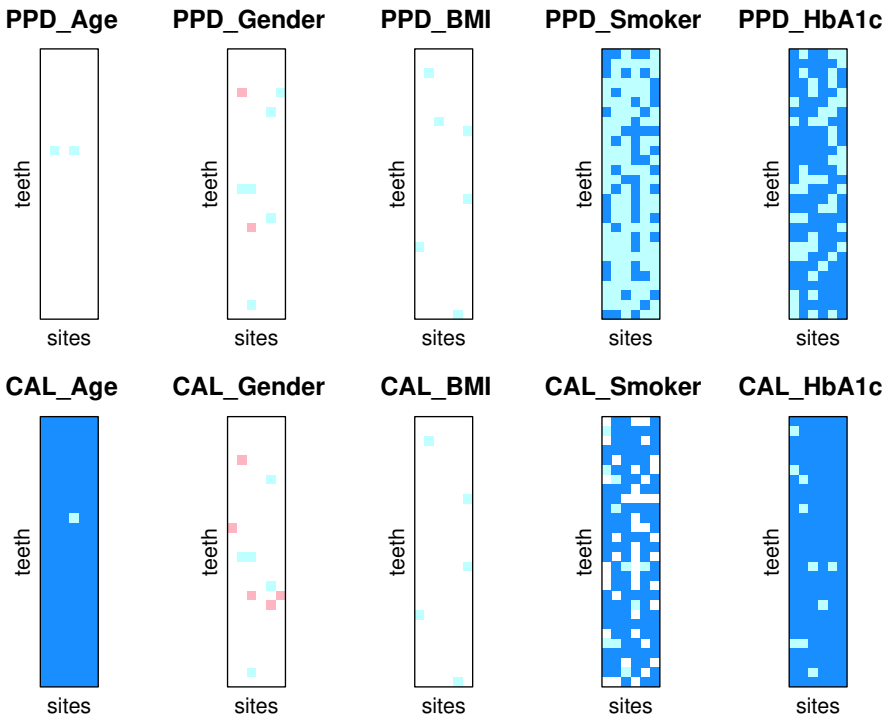

Figure 2.4: Fitting the BTN Model with sparsity prior on $\mathcal{B}$ to the GAAD data. Plotted are the D statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

types, respectively. While 'lingual' corresponds to the sites that are closest to the tongue, 'buccal' represents the diametrically opposite sites towards the cheek. Also, 'mesial/distal' represents sites that are closest to/away from the midline of the buccal cavity (the gap sites), while 'mid' denotes the sites lying at the midpoint on either (buccal/lingual) sides. From Table 2.3, we observe that Age exhibits the strongest effect on both PPD and CAL, irrespective of tooth-sites. This validates previous finding of the cumulative positive effect of age on PD [45]. Also, HbA1c exerts higher impact on the buccal sites compared to the lingual, for both PPD and CAL responses. The existing findings that patients with diabetes are susceptible to buccal area infections [40] confirm our results. For the assessments on tooth-types in Table 2.4, the 28 teeth (excluding the third molars) are classified as molars (#: 1-2, 13-16, 27-28), pre-molars (#: 3-4, 11-12, 17-18, 25-26),

canines (#: 5, 10, 19, 24), and incisors (#: 6-9, 20-23). Here also, the effect of Age is significantly prominent across all tooth-types. For HbA1c, the effect is the highest on the molars, followed by pre-molars, and other tooth types. For other covariates, such as gender, BMI and smoking, their effects on the molars are also the most prominent for both PPD and CAL responses. Previous analyses of the GAAD data revealed a significant proportion of diseased and missing molars [46], and the current finding strengthens this fact.

2.6 Conclusions

Our BSTN proposal enjoys several advantages over existing alternatives, such as (a) bypassing assumptions of either low-rank or sparsity structures on tensor coefficients, (b) allowing interpretation of marginal effects, (c) assessment of skewness levels and covariate effects separately on all marginal densities of tensor response, (d) straightforward handling of missing data under MAR, and (e) quantification of estimation uncertainty. We further demonstrate these advantages through well-designed simulation studies, and application to a real dataset on PD.

We further note that the tensor skewing shocks model of (2.2) extends the parametric version of "multiple skewing shocks" model of [7] to a tensor response, and are essentially different from the "single skewing shock" models

$$\mathcal{E}_i = |Z_{2i}|\mathcal{J} \times_K \mathbf{\Lambda}_K + \mathcal{Z}_{1i}, \tag{2.12}$$

with $\mathcal{J}$ being a tensor with all elements equal to 1. For the later, every element of the tensor error shares the same scalar skewing shock $|Z_{2i}|$, denoted by $e_{i,i_1\ldots i_K} = \lambda_{i_1}\cdots\lambda_{i_K}|z_{2i}| + z_{1i}$. The estimating equation method of [38] for multivariate skewed responses essentially uses this model, however, restricted only to a vector response. The corresponding within-subject association $\text{corr}(e_{i,i_1\ldots i_K}, e_{i,i'_1\ldots i'_K}) = \left\{\prod_{k=1}^K \rho_{i_k i'_k}\right\}\left[\left\{1+\lambda_{i_1}^2\cdots\lambda_{i_K}^2\left(1-\frac{2}{\pi}\right)\right\}\right]^{-1}$ is however, a function of the skewness parameters $\lambda_{i_1}\ldots\lambda_{i_K}$, because all elements of the tensor error share the common latent variable $|Z_{2i}|$. The Pearson's mode skewness of $e_{i,i_1\ldots i_K}$is $\lambda_{i_1}\cdots\lambda_{i_K}\sqrt{2/\pi}/[1+\lambda_{i_1}^2\cdots\lambda_{i_K}^2\{1-(2/\pi)\}]^{1/2}$ which is a function of $\lambda_{i1}\cdots\lambda_{i_K}$ in (2.12), while the marginal Pearson's first skewness coefficient is only a function of λ_{i_K} in (2.2). This hinders the prior specification of both skewness parameters λ_j and association parameters ρ_{i_k} based on the marginal prior opinion about the skewness and within subject association.

CHAPTER 3

A NEW CLASS OF SKEWED TENSOR DISTRIBUTIONS

3.1 Introduction

The long history of the class of elliptical distributions in multivariate data modeling possibly goes back to [15]. Early developments of elliptically symmetric distributions classes include [31], [9], [16], and [39]. See [18] for a comprehensive review of multivariate elliptical distributions and their examples. However, these works did not accommodate skewness in the multivariate responses. [1] introduced skewed elliptical distributions of linear form followed by a conditioning approach to introduce a related class of skewed elliptical distributions in [8]. Additional work by [2] proved close connection between these two forms of skew elliptical density classes. [48] introduced an alternative form of skew elliptical family containing multivariate skew normal and skew-t distributions as special cases. [24] extended this class to skewed matrix-variate elliptical distributions. However, extension of such a class of multivariate and matrix-variate skew-elliptical distributions to tensor responses of any higher order is still unavailable. In this chapter, we fill this gap between via proposing a new class of *skewed tensor elliptical distributions* (called STEL in short).

In addition to defining our new class of skewed tensor distributions called STEL and introducing its special cases such as skewed tensor normal and skewed tensor-t distributions, this article provides few more contributions. To the best of our knowledge, we are first to develop a theoretical framework for a class of skewed tensor elliptical distributions. Our STEL class is closed under marginalization to facilitate the interpretation of the covariate effects on any subset of tensor responses. Finally, we incorporate missing tensor responses within the Bayesian tensor regression regime. We also introduce a computationally feasible method to deal with regression of skewed tensor responses on a vector of covariates when only a sparse set of tensor components experience different regression effects than rest of the majority of tensor components. The remainder of the article is arranged as follows. In Section 3.2, we provide brief reviews of the class of multivariate elliptical distributions and the class of skewed multivariate elliptical distributions. In Section 3.3, we introduce our new

class of tensor elliptical distributions and its special cases. A new skewed elliptical class and its various sub-classes of distributions are presented in Section 3.4. In Section 3.5, we present the likelihood, hierarchical prior specification, and posterior distribution for a new Bayesian skewed tensor-t (BSTT) model for tensor response regression. Furthermore, we propose a tensor spike-and-slab Lasso prior. In Section 3.6, we demonstrate the practical advantages of our BSTT over alternatives via analysis of GAAD study. We conclude with a discussion and future research directions in Section 3.7. For ease of exposition, technical results and proofs are given in the Appendices.

3.2 Background

3.2.1 Multivariate Elliptical Class

Now, we define the class of multivariate elliptical distributions following Definition 2.2 in [18] and Section 2.1 in [48].

Definition 1. (Multivariate elliptical class) A random vector $\mathbf{y} \in \mathbb{R}^d$ has elliptical distribution with location $\boldsymbol{\mu} \in \mathbb{R}^d$ and scale-matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ if

$$\mathbf{y} \overset{d}{=} \boldsymbol{\mu} + \mathbf{A}\mathbf{x},\ \mathbf{x} \sim S_d(\phi), \tag{3.1}$$

where $\mathbf{x} \sim S_d(\phi)$ implies that $\mathbf{x}$ has a characteristic function of the form $\phi(\mathbf{t}^\intercal\mathbf{t})$, where $\phi(\cdot)$ is an univariate function, called the characteristic generator of the spherical distribution, and $\mathbf{A}^\intercal\mathbf{A} = \boldsymbol{\Sigma}$ for $\mathbf{A} \in \mathbb{R}^{k\times d}$ with $\text{rank}(\boldsymbol{\Sigma}) = k$. Here, the characteristic function and the density function (pdf) of $\mathbf{y}$ are respectively $\psi(\mathbf{t}) = \exp(i\mathbf{t}^\intercal\boldsymbol{\mu})\phi(\mathbf{t}^\intercal\boldsymbol{\Sigma}\mathbf{t})$ and

$$f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma};g^{(d)}) = \det(\boldsymbol{\Sigma})^{-\frac{1}{2}}g^{(d)}\{(\mathbf{y}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\},$$

where $g^{(d)} : \mathbb{R}^+ \to \mathbb{R}^+$ is defined as $g^{(d)} = \frac{\Gamma(d/2)}{\pi^{d/2}}\frac{g(u;d)}{\int_o^\infty r^{d/2-1}g(r;d)dr}$ for $g(\cdot;k) : \mathbb{R}^+ \to \mathbb{R}^+$. Here $r^{d/2-1}g(r;d)$ should be integrable over $[0,\infty)$ for $f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma};g^{(d)})$ to be a d-dimensional density, called the multivariate elliptical density denoted by $\mathbf{y} \sim EL(\boldsymbol{\mu},\boldsymbol{\Sigma};g^{(d)})$.

For the pdf in (3.1), we obtain the multivariate normal distribution when $g^{(d)}(u) = \frac{\exp(-u/2)}{(2\pi)^{d/2}}$, and the multivariate-$t$ distribution with the degree of freedom (df) ν when $g^{(d)}(u;\nu) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{d}{2}}}(1+\frac{u}{\nu})^{-\frac{\nu+d}{2}}$. We will define tensor elliptical class upon Definition 1 at section 3.3.

3.2.2 Skewed Multivariate Elliptical Class

The main goal of this section is to incorporate skewness in definition 1 using transformation method [48]. Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be a d-dimensional random vectors. We assume

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} \sim EL\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d \end{pmatrix}; g^{(2d)} \right),$$

where $\mathbf{0}$ denotes the null matrix and $\mathbf{I}_d$ is the $d \times d$ identity matrix. Now we use the transformation method to consider a skewed multivariate elliptical class of distributions,

$$\mathbf{y} = \boldsymbol{\Lambda}\mathbf{z}_2 + \mathbf{z}_1, \tag{3.2}$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix of skewness parameters. The skewed elliptical class would be developed with $\mathbf{y}|\mathbf{z}_2 > \mathbf{0}$, where $\mathbf{z}_2 > \mathbf{0}$ implying all elements in $\mathbf{z}_2$ are positive. When $\boldsymbol{\Lambda} = \mathbf{0}$, $\mathbf{y}$ becomes multivariate elliptical class (Definition 1). The equation (3.2) accompanied by $\mathbf{y}|\mathbf{z}_2 > \mathbf{0}$ allows skewness in multivariate response $\mathbf{y}$. The density of $\mathbf{y}|\mathbf{z}_2 > \mathbf{0}$ would be acquired using Theorem 1 of [48] as following definition 2.

Definition 2. (Skewed multivariate elliptical class) Let $\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\mu}$. Then the pdf of $\mathbf{y}|\mathbf{z}_2 > \mathbf{0}$ is

$$\begin{aligned}
&f\left(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}; g^{(d)}\right) = 2^d f_{\mathbf{y}}\left(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2; g^{(d)}\right) P(\mathbf{z}_2 > \mathbf{0}), \\
&\text{where} \quad \mathbf{z}_2 \sim EL\left(\boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}\tilde{\mathbf{y}}, \mathbf{I} - \boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}; g^{(d)}_{\tilde{\mathbf{y}}}\right), \\
&\text{and} \quad g^{(d)}_a(u) = \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}} \frac{g(a+u;2d)}{\int_0^\infty r^{\frac{d}{2}-1} g(a+r;2d)dr}, \quad a > 0, \\
&\text{and} \quad q(\tilde{\mathbf{y}}) = \tilde{\mathbf{y}}^\intercal (\boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2)^{-1}\tilde{\mathbf{y}}.
\end{aligned}$$

Then, we write $\mathbf{y} \sim SEL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}; g^{(d)})$, where "SEL" stands for skewed multivariate elliptical class of distributions.

We obtain a sub-class of definition 2, skewed multivariate normal distribution when $g^{(d)}(u) = (2\pi)^{-d/2}\exp(-u/2)$. Define $g(u; d, \nu) = (1 + \frac{u}{\nu})^{-(\nu+d)/2}$, where g depends on ν. Then we obtain $g^{(d)}(u; \nu)$ with straightforward calculation

$$g^{(d)}(u; \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{d}{2}}} g(u; d, \nu),$$

which produces skewed multivariate-t distribution. We extend Definition 2 to the general class of skewed tensor distributions.

3.3 Tensor Elliptical Class

We define the new class of tensor elliptical distributions and provide two special cases of the class.

Definition 3. (Tensor elliptical class) Let $\mathcal{M} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ and $\mathbf{\Sigma}_k \in \mathbb{R}^{d_k \times d_k}$ for $k = 1, \ldots, K$ be positive definite matrices. For any random tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ of order K, the density exists when the pdf of $\mathcal{Y}$ is given by

$$f(\mathcal{Y}|\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K; g^{(d_1,\ldots,d_K)}) = \prod_{k=1}^{K} \det(\mathbf{\Sigma}_k)^{-\frac{(d_1 \times \cdots \times d_K)}{(2d_k)}}$$
$$\times g^{(d_1,\ldots,d_K)}\left[\left\langle [\![\mathcal{Y} - \mathcal{M}; \mathbf{\Sigma}_1^{-\frac{1}{2}}, \ldots, \mathbf{\Sigma}_K^{-\frac{1}{2}}]\!], [\![\mathcal{Y} - \mathcal{M}; \mathbf{\Sigma}_1^{-\frac{1}{2}}, \ldots, \mathbf{\Sigma}_K^{-\frac{1}{2}}]\!] \right\rangle\right],$$

where the density generating function $g^{(d_1,\ldots,d_K)}(u)$ for $u \geq 0$ is defined as

$$g^{(d_1,\ldots,d_K)}(u) = \frac{\Gamma(\frac{d_1 \times \cdots \times d_K}{2})}{\pi^{\frac{(d_1 \times \cdots \times d_K)}{2}}} \frac{g(u; d_1, \ldots, d_K)}{\int_0^\infty r^{\frac{(d_1 \times \ldots \times d_K)}{2}-1} g(r; d_1, \ldots, d_K) dr},$$

for any $g(\cdot; d_1, \ldots, d_K) : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\int_0^\infty u^{\frac{(d_1 \times \ldots \times d_K)}{2}-1} g(u; d_1, \ldots$
$, d_K) dr < \infty$ exists. It is called the *tensor elliptical class* (TEL in short) of distributions denoted by

$$\mathcal{Y} \sim TEL\left(\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K; g^{(d_1,\ldots d_K)}\right).$$

The function $g(u; d_1, \ldots, d_K)$ is the kernel of and the rest of the terms in $g^{(d_1,\ldots,d_K)}(u)$ represent the normalizing constant of the density $f(\mathcal{Y}|\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K; g^{(d_1,\ldots,d_K)})$. The pdf of $\mathcal{Y}$ represents a broad class of tensor distributions. We introduce two special cases of TEL taking specific form for $g^{(d_1,\ldots,d_K)}$, tensor normal and tensor-t distributions.

Example 1. (Tensor Normal distribution) When $g(u; d_1, \ldots, d_K) = \exp(-u/2)$, We obtain $g^{(d_1,\ldots,d_K)}(u)$ $= \exp(-u/2)/(2\pi)^{(d_1 \times \cdots \times d_K)/2}$, and the pdf of $\mathcal{Y}$ becomes a tensor normal distribution (TN) denoted by $TN(\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K)$ with density

$$f(\mathcal{Y}|\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K; g^{(d_1,\ldots,d_K)}) = (2\pi)^{-\frac{(d_1 \times \cdots \times d_K)}{2}} \prod_{k=1}^{K} \det(\mathbf{\Sigma}_k)^{-\frac{(d_1 \times \cdots \times d_K)}{(2d_k)}}$$
$$\times \exp\left[-\frac{1}{2}\left\langle [\![\mathcal{Y} - \mathcal{M}; \mathbf{\Sigma}_1^{-\frac{1}{2}}, \ldots, \mathbf{\Sigma}_K^{-\frac{1}{2}}]\!], [\![\mathcal{Y} - \mathcal{M}; \mathbf{\Sigma}_1^{-\frac{1}{2}}, \ldots, \mathbf{\Sigma}_K^{-\frac{1}{2}}]\!] \right\rangle\right],$$

where $\det(\cdot)$ stands for determinant, and $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$ are positive definite symmetric matrices.

Example 2. (Tensor-t distribution) Let $g(u; d_1, \ldots, d_K, \nu) = (1 + \frac{u}{\nu})^{-(\nu+(d_1\times\cdots\times d_K))/2}$, where g depends on the df parameter, denoted by ν. A simple calculation produces

$$g^{(d_1,\ldots,d_K)}(u;\nu) = \frac{\Gamma\left(\frac{\nu+(d_1\times\cdots\times d_K)}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{(d_1\times\cdots\times d_K)}{2}}} g(u; d_1, \ldots, d_K, \nu),$$

where Γ is the gamma function. Thus, we obtain tensor-t distribution as follows.

$$f(\mathcal{Y}|\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K; g^{(d_1,\ldots,d_K)}) = \frac{\Gamma\left(\frac{\nu+(d_1\times\cdots\times d_K)}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{(d_1\times\cdots\times d_K)}{2}}} \prod_{k=1}^{K} \det(\mathbf{\Sigma}_k)^{-\frac{(d_1\times\cdots\times d_K)}{(2d_k)}}$$
$$\times \left[1 + \left\langle [\![\mathcal{Y} - \mathcal{M}; \mathbf{\Sigma}_1^{-\frac{1}{2}}, \ldots, \mathbf{\Sigma}_K^{-\frac{1}{2}}]\!], [\![\mathcal{Y} - \mathcal{M}; \mathbf{\Sigma}_1^{-\frac{1}{2}}, \ldots, \mathbf{\Sigma}_K^{-\frac{1}{2}}]\!] \right\rangle \Big/ \nu \right]^{-\frac{(\nu+(d_1\times\cdots\times d_K))}{2}},$$

which is the density of the $(d_1\times\cdots\times d_K)$-dimensional tensor-$t$ distribution denoted by $Tt(\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K, \nu)$. When $\nu = 1$, then $\mathcal{Y}$ becomes tensor Cauchy distribution. The marginal distribution is still tensor Cauchy distribution, however, the conditional distribution is tensor-t distribution. Even though the marginal means of the tensor Cauchy distribution do not exist for any dimensions, all the conditional means exist because the conditional distributions of the tensor Cauchy distribution follow tensor-t distributions with df larger than two.

3.4 Skewed Tensor Elliptical Class

In this section, we develop a new skewed tensor elliptical class using the transformation method introduced in [48]. Let $\mathcal{Z}_1 \sim TEL(0, \mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_K; g^{(d_1,\cdots,d_K)})$ and $\mathcal{Z}_2 \sim TEL(0, \mathbf{I}_{d_1}, \cdots, \mathbf{I}_{d_{K-1}}, \mathbf{D}^2_{\boldsymbol{\sigma}}; g^{(d_1,\cdots,d_K)})$ be two independent $(d_1 \times \cdots \times d_K)$-dimensional tensors, where $\mathbf{D}^2_{\boldsymbol{\sigma}} = \text{diag}(\sigma_1^2, \ldots, \sigma_{d_K}^2)$, g is a tensor density generator, 0 is the constant tensor that all elements are zero, $\mathbf{I}_d$ is a $d \times d$ identity matrix, and $\mathbf{\Sigma}_k$ is $d_k \times d_k$ positive definite matrix for $k = 1, \ldots, K$. We present a new skewed tensor elliptical class using

$$\begin{aligned} \mathcal{E} &= |\mathcal{Z}_2| \times_1 \mathbf{\Lambda}_1 \times_2 \mathbf{\Lambda}_2 \times_3 \cdots \times_K \mathbf{\Lambda}_K + \mathcal{Z}_1 \\ &= [\![|\mathcal{Z}_2|; \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K]\!] + \mathcal{Z}_1, \end{aligned} \tag{3.3}$$

where $\mathbf{\Lambda}_k = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kd_k}) \in \mathbb{R}^{d_k \times d_k}$ and skewenss parameters $\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kd_k})$ for $k = 1, \ldots, K$ modes. The distribution of tensor $\mathcal{E}$ in (3.3) allows different skewness levels for each mode. We consider the class of conditional distribution, $\mathcal{E}|\mathcal{Z}_2$, where $\mathcal{Z}_2 > 0$ implies all components in

$\mathcal{Z}_2$ are positive. When components of $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K$ are positive (negative), $\mathcal{E}$ in (3.3) becomes right (left) skewed tensor distributions. When $\boldsymbol{\lambda}_1 = \cdots = \boldsymbol{\lambda}_K = \mathbf{0}$, it retrieves to the tensor elliptical class (Definition 3).

Now we investigate theoretical properties of tensor elliptical class. The following two Theorems serve the bases to construct the skewed tensor elliptical class.

Theorem 2. (Linear Combination of Tensor Elliptical Variables). *Let $(d_1 \times \cdots \times d_K)$-dimensional tensor $\mathcal{Z}_2 \sim TEL(0, \boldsymbol{I}_{d_1}, \ldots, \boldsymbol{I}_{d_{K-1}}, \boldsymbol{D}^2_{\boldsymbol{\sigma}}; g^{(d_1,\ldots,d_K)})$. Define $\mathcal{M}$ be a $(d_1 \times \cdots \times d_K)$-dimensional constant tensor. Then, $\mathcal{M} + [\![\mathcal{Z}_2; \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K]\!]$ approximately follows $TEL(\mathcal{M}, \boldsymbol{\Lambda}_1^\intercal \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K^\intercal \boldsymbol{D}^2_{\boldsymbol{\sigma}} \boldsymbol{\Lambda}_K;$ $g^{(d_1,\ldots,d_K)})$.*

Theorem 2 indicates that any linear combination along each mode of tensor elliptically distributed variables is still tensor elliptical variables.

Theorem 3. (Conditional Density of Tensor Elliptical Variables). *Let the K-way tensors $\mathcal{Y}^*, \mathcal{M}^*,$ $\mathcal{Z}^* \in \mathbb{R}^{2d_1 \times \cdots \times 2d_K}$ be partitioned along their all modes simultaneously as $\mathcal{Y}^* = (\mathcal{Y}^{*(1)}, \mathcal{Y}^{*(2)})$, $\mathcal{M}^* = (\mathcal{M}^{*(1)}, \mathcal{M}^{*(2)})$, and $\mathcal{Z}^* = (\mathcal{Z}^{*(1)}, \mathcal{Z}^{*(2)})$ where $\mathcal{Y}^{*(\ell)}, \mathcal{M}^{*(\ell)}, \mathcal{Z}^{*(\ell)} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ for $\ell = 1, 2$. Partition $\boldsymbol{\Sigma}_k \in \mathbb{R}^{2d_k \times 2d_k}$ as $\begin{pmatrix} \boldsymbol{\Sigma}_{k,11} & \boldsymbol{\Sigma}_{k,12} \\ \boldsymbol{\Sigma}_{k,12} & \boldsymbol{\Sigma}_{k,22} \end{pmatrix}$, where $\boldsymbol{\Sigma}_{k,11} \in \mathbb{R}^{2d_{k,11} \times 2d_{k,11}}$, $\boldsymbol{\Sigma}_{k,12} \in \mathbb{R}^{2d_{k,12} \times 2d_{k,12}}$, $\boldsymbol{\Sigma}_{k,22} \in \mathbb{R}^{2d_{k,22} \times 2d_{k,22}}$, and $2(d_{k,11} + d_{k,12}) = 2(d_{k,12} + d_{k,22}) = 2d_k$ for $k = 1, \ldots, K$. Assume that $\mathcal{Z}^* \sim TEL(0, \boldsymbol{I}_{2d_1}, \ldots, \boldsymbol{I}_{2d_K};\ g^{(2d_1,\ldots,2d_K)})$, and $\mathcal{Y}^* \overset{d}{=} \mathcal{M}^* + [\![\mathcal{Z}^*; \boldsymbol{A}_1^\intercal, \ldots, \boldsymbol{A}_K^\intercal]\!]$, $\boldsymbol{\Sigma}_k = \boldsymbol{A}_k^\intercal \boldsymbol{A}_k$, then $\mathcal{Y}^{*(1)} | \mathcal{Y}^{*(2)} \sim TEL\left(\mathcal{M}^*_{1|2}, \boldsymbol{\Sigma}_{1,1|2}, \ldots, \boldsymbol{\Sigma}_{K,1|2};\ g^{(d_{1,1},\ldots,d_{K,1})}_{q(\mathcal{Y}^{*(2)})}\right)$, where $\mathcal{M}^*_{1|2} = \mathcal{M}^{*(1)} + [\![\mathcal{Y}^{*(2)} - \mathcal{M}^{*(2)};$ $\boldsymbol{\Sigma}_{1,12}\boldsymbol{\Sigma}^{-1}_{1,22}, \ldots, \boldsymbol{\Sigma}_{K,12}\boldsymbol{\Sigma}^{-1}_{K,22}]\!]$, $\boldsymbol{\Sigma}_{k,1|2} = \boldsymbol{\Sigma}_{k,11} - \boldsymbol{\Sigma}_{k,12}\boldsymbol{\Sigma}^{-1}_{k,22}\boldsymbol{\Sigma}_{k,21}$,*
$q(\mathcal{Y}^{*(2)}) = \left\langle \left[\!\left[\mathcal{Y}^{*(2)} - \mathcal{M}^{*(2)}; \boldsymbol{\Sigma}^{-\frac{1}{2}}_{1,22}, \ldots, \boldsymbol{\Sigma}^{-\frac{1}{2}}_{K,22} \right]\!\right], \left[\!\left[\mathcal{Y}^{*(2)} - \mathcal{M}^{*(2)}; \boldsymbol{\Sigma}^{-\frac{1}{2}}_{1,22}, \ldots, \boldsymbol{\Sigma}^{-\frac{1}{2}}_{K,22} \right]\!\right] \right\rangle$, and
$g_a^{(d_{1,1},\ldots,d_{K,1})}(u) = \frac{\Gamma(d_{1,1} \times \cdots \times d_{K,1}/2)}{\pi^{(d_{1,1} \times \cdots \times d_{K,1})/2}} \frac{g(a+u; d_1, \ldots, d_K)}{\int_0^\infty r^{(d_{1,1} \times \cdots \times d_{K,1})/2-1} dr}$ for some $r \geq 0$ which is independent of u.

Theorem 3 provides the straightforward calculation of conditional density of tensor elliptical variables. Also, it ensures we can easily obtain the likelihood contribution $f(\mathcal{Y}_{i,\text{obs}}|\Theta)$ of any observed tensor response $\mathcal{Y}_{i,\text{obs}}$ (possibly with missing elements) without integrating out the missing values numerically. When the tensor response is partitioned along its first mode (see, Corollary 1), it can be applied to the type-2 diabetic Gullah-speaking African-Americans (GAAD) study where any missing tooth i_1 of a patient has corresponding missing sites i_2 and biomarkers i_3 of $\mathcal{Y}_{i,i_1 i_2 i_3}$. We obtain conditional density $\mathcal{E}|\mathcal{Z}_2$ following the Theorem 4 using Theorems 2 and 3.

Theorem 4. (Skewed Tensor Elliptical Class). *The probability density function of* $\mathcal{E}|\mathcal{Z}_2 > \boldsymbol{0}$ *in (3.3) is given by*

$$\begin{aligned} & f(\mathcal{E}|0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K; g^{(d_1,\ldots,d_K)}) \\ & = 2^{(d_1 \times \cdots \times d_K)} \prod_{k=1}^{K-1} \det(\boldsymbol{Q}_k)^{-\frac{(d_1 \times \cdots \times d_{K-1})}{(2d_k)}} \det(\boldsymbol{Q}_K)^{-\frac{d_K}{2}} \\ & \times g^{(d_1,\ldots,d_K)} \left[\left\langle [\![\mathcal{E}; \boldsymbol{Q}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{Q}_K^{-\frac{1}{2}}]\!], [\![\mathcal{E}; \boldsymbol{Q}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{Q}_K^{-\frac{1}{2}}]\!] \right\rangle \right] \times P(\mathcal{Z}_2 > \boldsymbol{0}), \text{ where} \end{aligned} \tag{3.4}$$

$$\mathcal{Z}_2 \sim TEL\left([\![\mathcal{E}; \boldsymbol{\Lambda}_1 \boldsymbol{Q}_1^{-1}, \ldots, \boldsymbol{\Lambda}_K \boldsymbol{Q}_K^{-1}]\!] ; \boldsymbol{I}_{d_1} - \boldsymbol{\Lambda}_1 \boldsymbol{Q}_1^{-1} \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{D}_{\boldsymbol{\sigma}}^2 - \boldsymbol{\Lambda}_K \boldsymbol{Q}_K^{-1} \boldsymbol{\Lambda}_K; g_{q(\mathcal{E})}^{(d_1,\ldots,d_K)} \right), \tag{3.5}$$

where $\boldsymbol{Q}_k = \boldsymbol{\Sigma}_k + \boldsymbol{\Lambda}_k^{\intercal} \boldsymbol{\Lambda}_k$ *for* $k = 1, \ldots, K-1$, $\boldsymbol{Q}_K = \boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^{\intercal} \boldsymbol{D}_{\boldsymbol{\sigma}}^2 \boldsymbol{\Lambda}_K$,
$g_a^{(d_1,\ldots,d_K)}(u) = \frac{\Gamma(d_1 \times \cdots \times d_K/2)}{\pi^{(d_1 \times \cdots \times d_K)/2}} \frac{g(a+u; 2(d_1,\ldots,d_K))}{\int_0^\infty r^{(d_1 \times \cdots \times d_K)/2-1} g(a+u; 2(d_1,\ldots,d_K)) dr}$, $q(\mathcal{E}) = [\![\mathcal{E}; \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K]\!]$,
and $g_{q(\mathcal{E})}^{(d_1,\ldots,d_K)}$ *is free of* $q(\mathcal{E})$.

We denote $\mathcal{E}$ using the notation $\mathcal{E} \sim STEL\left(0; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K; g^{(d_1,\ldots,d_K)}\right)$, where $STEL$ stands for a class of skewed tensor elliptical distributions. We take specific forms in density generator function to provide special cases of $STEL$ in the subsection 3.4.1 and 3.4.2. One of the main contributions is that a skewed tensor elliptical class can be marginalized along a specific mode. We obtain the marginal distribution of $\mathcal{E}$ (subset of tensor response) as follows.

Proposition 1. (Marginal Density of Skewed Tensor Elliptical Variables) *Partition* $\mathcal{E} = \begin{pmatrix} \mathcal{E}^{(1)} \\ \mathcal{E}^{(2)} \end{pmatrix} \in \mathbb{R}^{2d_1 \times \cdots \times 2d_K}$, $0 = \begin{pmatrix} 0^{(1)} \\ 0^{(2)} \end{pmatrix}$, *and* $\mathcal{Z}_2 = \begin{pmatrix} \mathcal{Z}_2^{(1)} \\ \mathcal{Z}_2^{(2)} \end{pmatrix}$ *along* K*th mode, let* $d_K = 2$, $\boldsymbol{D}_{\boldsymbol{\sigma}}^2 = \text{diag}(\sigma_1^2, \sigma_2^2)$, $\boldsymbol{\Lambda}_K = \text{diag}(\lambda_1, \lambda_2)$, $\boldsymbol{R}_K = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, *and* $\boldsymbol{\Sigma}_K = \boldsymbol{D}_{\boldsymbol{\sigma}} \boldsymbol{R}_K \boldsymbol{D}_{\boldsymbol{\sigma}} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. *Here, particularly, we derive the first subset of components of* $\mathcal{E}$ *along* K*th mode. The marginal density of* $\mathcal{E}^{(1)}$ *would be*
$f\left(\mathcal{E}^{(1)}\right) = 2^{(d_1 \ldots d_{K-1})} \prod_{k=1}^{K-1} \det(\boldsymbol{Q}_k)^{-\frac{(d_1 \times \cdots \times d_{K-1})}{(2d_k)}} \det(\sigma_1^2 + \lambda_1^2)^{-\frac{1}{2}} \times g^{(d_1,\ldots,d_{K-1})}$
$\left[\left\langle [\![\mathcal{E}^{(1)}; \boldsymbol{Q}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{Q}_{K-1}^{-\frac{1}{2}}, (\sigma_1^2+\lambda_1^2)^{-\frac{1}{2}}]\!], [\![\mathcal{E}^{(1)}; \boldsymbol{Q}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{Q}_{K-1}^{-\frac{1}{2}}, (\sigma_1^2+\lambda_1^2)^{-\frac{1}{2}}]\!] \right\rangle \right] \times P(\mathcal{Z}_2^{(1)} > 0)$, *where*
$\mathcal{Z}_2^{(1)} \sim TEL\left(\lambda_1(\sigma_1^2 + \lambda_1^2)^{-1} \mathcal{E}^{(1)}; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{K-1}, \sigma_1^2 - \lambda_1(\sigma_1^2 + \lambda_1^2)^{-1} \lambda_1; g^{(d_1,\ldots,d_{K-1})} \right)$.

The marginal density ($\mathcal{E}^{(1)}$) in Proposition 1 has the same form of the density ($\mathcal{E}$) in Theorem 4. That is, the distribution of any subset of the skewed tensor response is within the same skewed

tensor distribution of class of the original tensor. Thus, the class of skewed tensor elliptical variables can be marginalized. In GAAD study, for instance, the density of subset of tensor response is a function of relevant parameters such as skewness, and variability of each biomarker.

3.4.1 Skewed Tensor Normal Distribution

Let $g^{(d_1,\ldots,d_K)}(u) = \exp(-u/2)/(2\pi)^{(d_1\times\cdots\times d_K)/2}$ in (3.4) - (3.5), then the density of skewed tensor elliptical distribution of $\mathcal{E}$ becomes skewed tensor normal distribution as follows.

$$
\begin{aligned}
&f\big(\mathcal{E}|0,\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K;\mathbf{\Lambda}_1,\ldots,\mathbf{\Lambda}_K\big) = (2\pi)^{-\frac{(d_1\times\cdots\times d_K)}{2}}\prod_{k=1}^{K-1}\det(\mathbf{Q}_k)^{-\frac{(d_1\times\cdots\times d_{K-1})}{(2d_k)}}\det(\mathbf{Q}_K)^{-\frac{d_K}{2}}\\
&\times\exp\left[-\frac{1}{2}\left\langle[\![\mathcal{E};\mathbf{Q}_1^{-\frac{1}{2}},\ldots,\mathbf{Q}_K^{-\frac{1}{2}}]\!],[\![\mathcal{E};\mathbf{Q}_1^{-\frac{1}{2}},\ldots,\mathbf{Q}_K^{-\frac{1}{2}}]\!]\right\rangle\right]\times P(\mathcal{Z}_2>0),\\
&\text{where } \mathcal{Z}_2\sim TN\bigg([\![\mathcal{E};\mathbf{\Lambda}_1\mathbf{Q}_1^{-1},\ldots,\mathbf{\Lambda}_K\mathbf{Q}_K^{-1}]\!];\mathbf{I}_{d_1}-\mathbf{\Lambda}_1\mathbf{Q}_1^{-1}\mathbf{\Lambda}_1,\ldots,\mathbf{D}^2_{\boldsymbol{\sigma}}-\mathbf{\Lambda}_K\mathbf{Q}_K^{-1}\mathbf{\Lambda}_K\bigg).
\end{aligned}
$$

We denote the above distribution by $STN(0,\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K;\mathbf{\Lambda}_1,\ldots,\mathbf{\Lambda}_K)$, where STN represents skewed tensor normal distribution and $\mathbf{\Sigma}_k$ takes equicorrelation matrix for $k=1,\ldots,K$. The skewed tensor normal distribution is a subclass of skewed tensor elliptical class. It is usually hard to compute $P(\mathcal{Z}_2>0|\mathcal{E})$. Thus, we construct hierarchical model of $f\big(\mathcal{Y}\big||\mathcal{Z}_2|\big)$ and $f(|\mathcal{Z}_2|)$ to do inference in a Bayesian context (see, (3.17) in section 3.5).

Moments. We derive the moment generating function (mgf) to obtain first two moments of the skewed tensor normal distribution, $\mathcal{Y}\sim TSN(\mathcal{M},\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K;\mathbf{\Lambda}_1,\ldots,\mathbf{\Lambda}_K)$. The mgf is given by

$$
\begin{aligned}
M_{\mathcal{Y}}(\mathcal{T}) &= 2^{(d_1\times\cdots\times d_K)}P(\mathcal{Z}\le[\![\mathcal{T};\mathbf{\Lambda}_1,\ldots,\mathbf{\Lambda}_K]\!])\\
&\quad\times\exp\left\{\frac{1}{2}\left\langle[\![\mathcal{T};\mathbf{G}_1^{\frac{1}{2}},\ldots,\mathbf{G}_K^{\frac{1}{2}}]\!],[\![\mathcal{T};\mathbf{G}_1^{\frac{1}{2}},\ldots,\mathbf{G}_K^{\frac{1}{2}}]\!]\right\rangle+\langle\mathcal{M},\mathcal{T}\rangle\right\},
\end{aligned}
\tag{3.6}
$$

where $\mathcal{Z}\sim TN(0,\mathbf{I}_{d_1},\ldots,\mathbf{I}_{d_K})$, and $\mathbf{G}_1=(\mathbf{\Sigma}_1+\mathbf{\Lambda}_1^2)^{-1},\ldots,\mathbf{G}_K=(\mathbf{\Sigma}_K+\mathbf{\Lambda}_K^2)^{-1}$. The details of derivation mgf (3.6), are given by Appendices. The first moment of $\mathcal{Y}$ is $E(\mathcal{Y})=\mathcal{M}+[\![\sqrt{\frac{2}{\pi}}\mathcal{J};\mathbf{\Lambda}_1,\ldots,\mathbf{\Lambda}_K\mathbf{D}^2_{\boldsymbol{\sigma}}]\!]$, where $\mathcal{J}$ denotes K-way identity tensor. The second moment, $\text{cov}(\text{vec}(\mathcal{Y}))=\mathbf{D}^2_\sigma\{\mathbf{\Sigma}_K+(1-\frac{2}{\pi})\mathbf{\Lambda}_K^2\}\otimes\cdots\otimes\{\mathbf{\Sigma}_1+(1-\frac{2}{\pi})\mathbf{\Lambda}_1^2\}$, where $\mathbf{\Lambda}_k=\text{diag}(\lambda_{k1},\ldots,\lambda_{kd_k})$ for $k=1,\ldots,K$. Note that skewness and correlation structure do not influence each other. The value of skewness changes correlation, but cannot affect the structure.

Skewness and Correlation. The model in (3.3) assumes that each element $y_{i_1\ldots i_K}$ of tensor response $\mathcal{Y}$ has its independent skewing shock $|z_{2,i_1\ldots i_K}|$. So, $\mathrm{E}(e_{i_1,\ldots,i_K}) = \prod_{k=1}^{K}\lambda_{ki_k}\mathrm{E}(|z_{2,i_1,\ldots,i_K}|)+\mathrm{E}(z_{1,i_1,\ldots,i_K}) = \sigma_{Ki_k}\prod_{k=1}^{K}\lambda_{ki_k}\sqrt{\frac{2}{\pi}}$ and variance $\mathrm{Var}(e_{i_1\ldots i_K}) = \prod_{k=1}^{K}\lambda_{ki_k}^2\mathrm{Var}(|z_{2,i_1,\ldots,i_K}|)+\mathrm{Var}(z_{1,i_1,\ldots,i_K}) = \sigma_{Ki_k}^2\prod_{k=1}^{K}\lambda_{ki_k}^2(1-\frac{2}{\pi})+\sigma_{Ki_k}^2 = \sigma_{Ki_k}^2[1+\prod_{k=1}^{K}\lambda_{ki_k}^2\{1-(\frac{2}{\pi})\}]$ for $k = 1,\ldots,K$. Thus, the Pearson's first skewness coefficient [44] of $e_{i_1\ldots i_K}$ in (3.3) would be calculated by $\frac{\{\mathrm{E}(e_{i_1,\ldots,i_K})-\mathrm{mode}(e_{i_1,\ldots,i_K})\}}{\sqrt{\mathrm{Var}(e_{i_1,\ldots,i_K})}} = \frac{\prod_{k=1}^{K}\lambda_{ki_k}\sqrt{\frac{2}{\pi}}}{\sqrt{[1+\prod_{k=1}^{K}\lambda_{ki_k}^2\{1-(\frac{2}{\pi})\}]}}$, where the $\mathrm{mode}(e_{i_1,\ldots,i_K})$ is zero. Here, the mode is the most frequent value of $e_{i_1,\ldots,i_K}$. Therefore, the Pearson's first skewness coefficient of our response distribution is a function of $\prod_{k=1}^{K}\lambda_{i_K}$, but does not depend on σ_{i_K} and $\boldsymbol{\Sigma}_k$.

The correlation between any pair of responses within a subject is

$$\mathrm{corr}(e_{i_1\ldots i_K}, e_{i'_1\ldots i'_K}) \;=\; \frac{\prod_{k=1}^{K}\rho_{i_k i'_k}}{\sqrt{\left[\left\{1+\prod_{k=1}^{K}\lambda_{ki_k}^2\left(1-\frac{2}{\pi}\right)\right\}\left\{1+\prod_{k=1}^{K}\lambda_{ki'_k}^2\left(1-\frac{2}{\pi}\right)\right\}\right]}}. \tag{3.7}$$

The correlation coefficient in (3.7) is a function of skewness parameters $\boldsymbol{\lambda}_j = (\lambda_{j1},\ldots,\lambda_{jd_j})$ for $j = 1,\ldots,K$, and correlation parameters $\rho_{i_k i'_k}$ for $k = 1,\ldots,K$, but it is not depend on the scale parameters $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_{d_K})$.

3.4.2 Skewed Tensor-t Distribution

Let $g^{(d_1,\ldots,d_K)}(u;\nu) = (1+\frac{u}{\nu})^{-\frac{(\nu+2(d_1\times\cdots\times d_K))}{2}}$. We need a marginal density from (3.4), and a cumulative conditional density of (3.5) to define skewed tensor-t distribution. For $(d_1\times\cdots\times d_K)$-dimensional marginal density, we have

$$g^{(d_1,\ldots,d_K)}(u) = \frac{\Gamma(d_1\times\cdots\times d_K)}{\pi^{\frac{(d_1\times\cdots\times d_K)}{2}}}\frac{g(u;d_1,\ldots,d_K,\nu)}{\int_0^\infty r^{\frac{(d_1\times\cdots\times d_K)}{2}-1}g(r;d_1,\ldots,d_K,\nu)dr}, \tag{3.8}$$

following Lemma 10 in the Appendices. Thus, the marginal density in (3.4) taking (3.8) follows tensor-t distribution,

$$g^{(d_1,\ldots,d_K)}\left[\left\langle[\![\mathcal{E};\mathbf{Q}_1^{-\frac{1}{2}},\ldots,\mathbf{Q}_K^{-\frac{1}{2}}]\!],[\![\mathcal{E};\mathbf{Q}_1^{-\frac{1}{2}},\ldots,\mathbf{Q}_K^{-\frac{1}{2}}]\!]\right\rangle\right] \sim Tt(0,\mathbf{Q}_1,\ldots,\mathbf{Q}_K,\nu).$$

Now, the cumulative conditional density for $\mathcal{Z}_2$ would be obtained via (3.5) (see, details in Lemma 11 in Appendices).

$$Tt\left(\mathcal{Z}_2\middle|[\![\mathcal{E};\boldsymbol{\Lambda}_1\mathbf{Q}_1^{-1},\ldots,\boldsymbol{\Lambda}_K\mathbf{Q}_K^{-1}]\!],\frac{\nu+\mathcal{Y}^*}{\nu^*}\{\mathbf{I}_{d_1}-\boldsymbol{\Lambda}_1\mathbf{Q}_1^{-1}\boldsymbol{\Lambda}_1,\ldots,\mathbf{D}_{\boldsymbol{\sigma}}^2-\boldsymbol{\Lambda}_K\mathbf{Q}_K^{-1}\boldsymbol{\Lambda}_K\},\nu^*\right), \tag{3.9}$$

where $\mathcal{Y}^* = \langle [\![\mathcal{Y} - \mathcal{M}; \mathbf{Q}_1^{-\frac{1}{2}}, \ldots, \mathbf{Q}_K^{-\frac{1}{2}}]\!], [\![\mathcal{Y} - \mathcal{M}; \mathbf{Q}_1^{-\frac{1}{2}}, \ldots, \mathbf{Q}_K^{-\frac{1}{2}}]\!] \rangle$, and $\nu^* = \nu + (d_1 \times \cdots \times d_K)$. Two terms are added on (3.9) which are not included in skewed tensor normal distribution. First, interestingly, the degrees of freedom of the conditional distribution increases to $(\nu + (d_1 \times \cdots \times d_K))$. The higher dimensions of tensor responses we have, the less heavy-tailedness we have. Thus, one could anticipate that the best choice of ν should close to two (but larger than two). Second, $(\nu + \mathcal{Y}^*)/\nu^*$ controls scale of each mode of conditional density. As each dimension of tensor response goes higher, we obtain smaller variability in each mode.

Theorem 5 derives skewed tensor-t distribution of $\text{vec}(\mathcal{Y})$ when one may build hierarchical model. Two levels of setup will be useful for Bayesian inference in section 3.5.1.

Theorem 5. *Let* $\mathcal{Y} = \mathcal{M} + [\![\mathcal{Z}_2; \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K]\!]$. *Two level of hierarchical model is assumed to be* $\mathcal{Y}|\mathcal{Z}_2 \sim TN(\mathcal{M} + [\![\mathcal{Z}_2; \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K]\!], \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K/\delta)$, *and* $\mathcal{Z}_2 \sim TN(0, \boldsymbol{I}_{d_1}, \ldots, \boldsymbol{I}_{d_{K-1}}, \boldsymbol{D}_{\boldsymbol{\sigma}}^2/\delta)$. *Then,* $\mathcal{Y} \sim STT(\mathcal{M}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K; \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K, \nu)$ *if* $\text{vec}(\mathcal{Y})$ *has the pdf*

$$
\begin{aligned}
&f(\text{vec}(\mathcal{Y})|\text{vec}(\mathcal{M}), \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K, \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K, \nu) \\
&= 2^{(d_1 \ldots d_K)} t_{\prod_{k=1}^K d_k} \left(\text{vec}(\mathcal{Y})|\text{vec}(\mathcal{M}), \mathbf{\Sigma}^* + (\mathbf{\Lambda}_K^\intercal \boldsymbol{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1), \nu\right) \\
&\times T_{\prod_{k=1}^K d_k} \left(\boldsymbol{z}_2^* \middle| \boldsymbol{D}_{\boldsymbol{\sigma}}^2 \otimes \boldsymbol{I}_{d_1 \ldots d_{K-1}} - \mathbf{\Lambda}^* \{\mathbf{\Sigma}^* + (\mathbf{\Lambda}_K^\intercal \boldsymbol{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1} \mathbf{\Lambda}^*, \nu + \prod_{k=1}^K d_k \right), \\
&\text{where } \boldsymbol{z}_2^* = [\mathbf{\Lambda}^* \{\mathbf{\Sigma}^* + (\mathbf{\Lambda}_K^\intercal \boldsymbol{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1}](\text{vec}(\mathcal{Y}) - \text{vec}(\mathcal{M})) \\
&\times \sqrt{\frac{\nu + \prod_{k=1}^K d_k}{\nu + (\text{vec}(\mathcal{Y}) - \text{vec}(\mathcal{M}))^\intercal \{\mathbf{\Sigma}^* + (\mathbf{\Lambda}_K^\intercal \boldsymbol{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1} (\text{vec}(\mathcal{Y}) - \text{vec}(\mathcal{M}))}} \\
&\text{and we denote } T_{\prod_{k=1}^K d_k}(\cdot|\mathbf{\Sigma}^*, \nu) \text{ as the cumulative distribution function (henceforth, cdf)} \\
&\text{of } t_{\prod_{k=1}^K d_k}(0, \mathbf{\Sigma}^*, \nu), \text{ and } t_{\prod_{k=1}^K d_k} \text{ represents } \left(\prod_{k=1}^K d_k\right)\text{-variate } t \text{ distribution,}
\end{aligned}
$$

where $\mathbf{\Lambda}^* = \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1$, $\mathbf{\Sigma}^* = \mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1$, *and STT stands for skewend tensor-t distribution.*

Moments. Since skewed tensor-t distribution is a scale mixture of skewed tensor normal distribution, we can obtain by using mgf of skewed tensor normal distribution.

$$
\begin{aligned}
M_{\mathcal{Y}}(\mathcal{T}) = 2^{(d_1 \times \cdots \times d_K)} &\int_0^\infty \exp\left\{ \frac{1}{2\delta} \left\langle [\![\mathcal{T}; \mathbf{G}_1^{\frac{1}{2}}, \ldots, \mathbf{G}_K^{\frac{1}{2}}]\!], [\![\mathcal{T}; \mathbf{G}_1^{\frac{1}{2}}, \ldots, \mathbf{G}_K^{\frac{1}{2}}]\!] \right\rangle \right\} \\
&\times P(\mathcal{Z} \leq \delta^{-\frac{1}{2}} [\![\mathcal{T}; \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K]\!]) d\Psi(\delta),
\end{aligned} \tag{3.10}
$$

where $\mathbf{G}_1 = (\mathbf{\Sigma}_1 + \mathbf{\Lambda}_1^2)^{-1}, \ldots, \mathbf{G}_K = (\mathbf{\Sigma}_K + \mathbf{\Lambda}_K^2)^{-1}$, $\mathcal{Z} \sim TN(0, \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_K})$, $\Psi(\delta)$ represents the cdf of $\Gamma(\frac{\nu}{2}, \frac{\nu}{2})$. Following the properties of the multivariate skew-t (MST) distribution, the mean and the covariance of $\mathcal{Y}$ are given by $E(\mathcal{Y}) = \mathcal{M} + [\![\sqrt{\frac{\nu}{\pi}}\frac{\Gamma(\nu-1/2)}{\Gamma(\nu/2)}; \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K \mathbf{D}_\sigma^2]\!]$, and $\mathrm{Cov}(\mathcal{Y}) = \mathbf{D}_\sigma^2[\{\mathbf{\Sigma}_K + \mathbf{\Lambda}_K^2\}\frac{\nu}{\nu-2} - \frac{\nu}{\pi}\{\frac{\Gamma(\nu-1/2)}{\Gamma(\nu/2)}\}^2\mathbf{\Lambda}_1^2] \otimes \cdots \otimes [\{\mathbf{\Sigma}_1 + \mathbf{\Lambda}_1^2\}\frac{\nu}{\nu-2} - \frac{\nu}{\pi}\{\frac{\Gamma(\nu-1/2)}{\Gamma(\nu/2)}\}^2\mathbf{\Lambda}_1^2]$, where $\nu > 2$. When $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_{K-1} = 1$, then STT defined above retrieve to the MVT of [48].

Skewness and Correlation. Since $|z_{2,i_1,\ldots,i_K}|$ follows half-t distribution, $\mathrm{E}(e_{i_1,\ldots,i_K}) = \prod_{k=1}^K \lambda_{ki_k}\mathrm{E}(|z_{2,i_1,\ldots,i_K}|) + \mathrm{E}(z_{1,i_1,\ldots,i_K}) = 2\sigma_{Ki_k}\prod_{k=1}^K \lambda_{ki_k}\sqrt{\frac{\nu}{\pi}}\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})(\nu-1)}$ for $\nu > 1$ and variance $\mathrm{Var}(e_{i_1\ldots i_K}) = \prod_{k=1}^K \lambda_{ki_k}^2\mathrm{Var}(|z_{2,i_1,\ldots,i_K}|) + \mathrm{Var}(z_{1,i_1,\ldots,i_K}) = \sigma_{Ki_k}^2\prod_{k=1}^K \lambda_{ki_k}^2\{\frac{\nu}{\nu-2} - \frac{4\nu}{\pi(\nu-1)^2}(\frac{\Gamma(\nu+1/2)}{\Gamma(\nu/2)})^2\} + \sigma_{Ki_k}^2 = \sigma_{Ki_k}^2[1 + \prod_{k=1}^K \lambda_{ki_k}^2\{\frac{\nu}{\nu-2} - \frac{4\nu}{\pi(\nu-1)^2}(\frac{\Gamma(\nu+1/2)}{\Gamma(\nu/2)})^2\}]$ for $\nu > 2$, and $k = 1, \ldots, K$. So, the Pearson's first skewness coefficient of $e_{i_1\ldots i_K}$ in (3.3) is

$$\frac{2\prod_{k=1}^K \lambda_{ki_k}\sqrt{\frac{\nu}{\pi}}\frac{\Gamma(\nu+1/2)}{\Gamma(\nu/2)(\nu-1)}}{\sqrt{1+\prod_{k=1}^K \lambda_{ki_k}^2}\left\{\sqrt{\frac{\nu}{\nu-2}} - \sqrt{\frac{2\nu}{\pi(\nu-1)}}\left(\frac{\Gamma(\nu+1/2)}{\Gamma(\nu/2)}\right)\right\}} \quad \text{for } \nu > 2.$$

Unlike the skewness coefficient of skewed tensor normal distribution, the Pearson's first skewness coefficient for skewed tensor-t distribution depends on degrees of freedom and $\prod_{k=1}^K \lambda_{ki_k}^2$, but not on σ_{Ki_K}.

The correlation between two different response within a patient is

$$\mathrm{corr}(e_{i_1\ldots i_K}, e_{i'_1\ldots i'_K}) = \frac{\prod_{k=1}^K \rho_{i_k i'_k}}{\sqrt{\left(1+\prod_{k=1}^K \lambda_{ki_k}^2\right)\left(1+\prod_{k=1}^K \lambda_{ki'_k}^2\right)}\left\{\sqrt{\frac{\nu}{\nu-2}} - \sqrt{\frac{2\nu}{\pi(\nu-1)}}\left(\frac{\Gamma(\nu+1/2)}{\Gamma(\nu/2)}\right)\right\}^2}. \tag{3.11}$$

The correlation in (3.11) depends on degrees of freedom, skewenss parameters and correlation parameters, but it is not a function of scale parameters $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{d_K})$.

3.4.3 Skewed Tensor Response Regression Model

In this subsection, we introduce a new skewed tensor-t regression model. Note that it is a scale mixture of a skewed tensor normal regression model. For a K-way tensor response $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ and a p-dimensional vector of covariates $\mathbf{x}_i$, a tensor regression model is

$$\mathcal{Y}_i = \mathcal{B}\bar{\times}_{(K+1)}\mathbf{x}_i + \mathcal{E}_i, \quad \text{for} \quad i = 1, \ldots, n, \tag{3.12}$$

where $\mathcal{B} \in \mathbb{R}^{d_1 \times \cdots \times d_K \times p}$ is an $(K+1)$th order unknown tensor of regression coefficients, $\bar{\times}_{(K+1)}$ is the $(K+1)$-mode vector product, and $\mathcal{E}_i \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a Kth order tensor of error. We consider

the special case of (3.3) by assuming skewness solely exists on Kth mode on $\mathcal{E}_i$ as follows,

$$\begin{aligned}\mathcal{E}_i &= |\mathcal{Z}_{2i}| \times_1 \mathbf{R}_1^{\frac{1}{2}} \cdots \times_{K-1} \mathbf{R}_{K-1}^{\frac{1}{2}} \times_K \mathbf{\Lambda}_K + \mathcal{Z}_{1i}, \\ &= \left[\!\!\left[|\mathcal{Z}_{2i}|; \mathbf{R}_1^{\frac{1}{2}}, \ldots, \mathbf{R}_{K-1}^{\frac{1}{2}}, \mathbf{\Lambda}_K \right]\!\!\right] + \mathcal{Z}_{1i},\end{aligned} \tag{3.13}$$

where $\mathbf{\Lambda}_K = \text{diag}(\lambda_1, \ldots, \lambda_{d_K}) \in \mathbb{R}^{d_K \times d_K}$ denotes diagonal matrix of skewness parameters. In (3.13), d_K skewness parameters are only required to be estimated, while (3.3) needs to estimate $\sum_{k=1}^{K} d_k$ parameters. Thus, the assumption of skewness in (3.13) is reasonable in real applications compared to (3.3). We use $|\mathbf{A}|$ to denote each component of $|\mathbf{A}|$ is the absolute value of the corresponding component of $\mathbf{A}$. In (3.13), $|\mathcal{Z}_{2i}|$ denotes the tensor skewing shock, where $\mathcal{Z}_{2i} \sim TN(\mathbf{0}; \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}_{\boldsymbol{\sigma}}^2/\delta_i)$. It is assumed to be independent of $\mathcal{Z}_{1i} \sim TN(\mathbf{0}; \mathbf{R}_1, \ldots, \mathbf{R}_{K-1},$ $\mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}}/\delta_i)$, where $\mathbf{R}_k > 0$ is a correlation matrix for $k = 1, \ldots, K$, $\mathbf{D}_{\boldsymbol{\sigma}} = \text{diag}(\sigma_1, \ldots, \sigma_{d_K})$ with $\sigma_j > 0$ for $j = 1, \cdots, d_K$. A latent variable δ_i has Gamma distribution, $Ga(\nu/2, \nu/2)$, and it scales covariance matrix on Kth mode. When $\mathbf{\Lambda}_K = \mathbf{0}$, then $\mathcal{E}_i$ of (3.13) retrieve to $\mathcal{Z}_{1i}$ following a tensor normal distribution. We assume separable covariance structure, $\text{cov}\{\text{vec}(\mathcal{Z}_{1i})\} = (\mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}}/\delta_i) \otimes \mathbf{R}_{K-1} \otimes \cdots \otimes \mathbf{R}_1$ ([26, 27, 33]). This separable covariance structure alleviates computational burden for the Bayesian inference. We denote $\mathcal{E}_i$ in (3.13) having skewed tensor normal distribution $\mathcal{E}_i \sim STN(\mathbf{0}, \mathbf{R}_1, \ldots, \mathbf{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$ as defined in section 3.4.1. The $\mathbf{R}_1, \ldots, \mathbf{R}_K$ are equicorrelation matrices to ensure the identifiability of the model parameters in (3.13).

As we shown in Proposition 1, the model of (3.12)-(3.13) is closed under marginalization to facilitate the interpretation of the covariate effect, skewness, and scale parameters on any subset of tensor responses. Particularly, in our motivating GAAD study, we can interpret covariate effect, skewness level, and variability of each biomarker.

The model (3.12)-(3.13) would be vectorized,

$$\text{vec}(\mathcal{Y}_i) = \mathbf{B}_{(K+1)}^{\intercal}\mathbf{x}_i + \left(\mathbf{\Lambda}_K \otimes \mathbf{R}_{K-1}^{\frac{1}{2}} \otimes \cdots \otimes \mathbf{R}_1^{\frac{1}{2}}\right) \text{vec}(|\mathcal{Z}_{2i}|) + \text{vec}(\mathcal{Z}_{1i}), \quad \text{for} \quad i = 1, \ldots, n, \tag{3.14}$$

where $\text{vec}(\mathcal{Y}_i) \in \mathbb{R}^{\prod_{k=1}^{K} d_k}$ is a multivariate response, $\mathbf{B}_{(K+1)} \in \mathbb{R}^{p \times \prod_{k=1}^{K} d_k}$ is the mode-$(K+1)$ matricization (as defined at the section 1.3) of the tensor coefficient $\mathcal{B}$ in (3.12). We remark that each column of $\mathbf{B}_{(K+1)}$ in (3.14) is a vector of coefficient describing the linear relationship between the individual elements of $\mathcal{Y}_i$ and the covariate $\mathbf{x}_i$.

When $K = 1$, the skewed tensor error in (3.13) reduces to the d_K-variate error model,

$$\mathbf{e}_i = \mathbf{\Lambda}_K|\mathbf{z}_{2i}| + \mathbf{z}_{1i}, \tag{3.15}$$

where $\mathbf{z}_{1i} \sim N_{d_K}(\mathbf{0}, \mathbf{D}_{\boldsymbol{\sigma}}\mathbf{R}_K\mathbf{D}_{\boldsymbol{\sigma}})$, $\mathbf{z}_{2i} \sim N_{d_K}(\mathbf{0}, \mathbf{D}^2_{\boldsymbol{\sigma}})$. We denote $\mathbf{e}_i$ as Multivariate Skew Normal distribution, $MSN(\mathbf{0}, \mathbf{R}_K; \boldsymbol{\sigma}, \boldsymbol{\Lambda}_K)$ of [7]. As the multivariate model of (3.15), the skewness of tensor error $\mathcal{E}_i$ of (3.13) is modeled via latent tensor "skewing shocks" $|\mathcal{Z}_{2i}|$ and the skewness parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_K})$ of $\boldsymbol{\Lambda}_K$. When λ_j is positive (or negative), the corresponding marginal density of $y_{i_1,\ldots,i_{K-1},j}$ of tensor response $\mathcal{Y}$ is skewed to the right (left).

3.5 Bayesian Inference

We develop a Bayesian inferential framework for a tensor response with missing data. Also, we introduce two types of prior for tensor regression coefficient; 1) Tensor normal prior and 2) Tensor Spike-and-Slab Lasso (TSSL) prior.

3.5.1 Likelihood, Hierarchical Prior and Posterior

Given the observed tensor response data $\mathcal{Y}_i$ from $i = 1, \ldots, n$ subjects, the likelihood function for the parameters $\Theta = (\mathcal{B}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \boldsymbol{\sigma}, \nu)$ is given by

$$\begin{aligned} L(\Theta|\mathcal{Y}) \propto \prod_{i=1}^{n} \int & f_1(\mathcal{Y}_i - \mathcal{B}\bar{\times}_{(K+1)}\mathbf{x}_i - |\mathcal{Z}_{2i}| \times_K \boldsymbol{\Lambda}_K \big| \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{K-1}, \mathbf{D}_{\boldsymbol{\sigma}}\boldsymbol{\Sigma}_K\mathbf{D}_{\boldsymbol{\sigma}}/\delta_i) \\ & \times f_2(\mathcal{Z}_{2i} \big| \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}^2_{\boldsymbol{\sigma}}/\delta_i) d\mathcal{Z}_{2i} \times f_3(\delta_i/\nu), \end{aligned} \tag{3.16}$$

where f_1 represents the density of $TN(\mathbf{0}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{K-1}, \mathbf{D}_{\boldsymbol{\sigma}}\boldsymbol{\Sigma}_K\mathbf{D}_{\boldsymbol{\sigma}}/\delta_i)$ of $\mathcal{Z}_{1i}$ in (3.13), f_2 represents the density of a latent tensor, $\mathcal{Z}_{2i} \sim TN(\mathbf{0}, \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}^2_{\boldsymbol{\sigma}}/\delta_i)$ in (3.13). We introduce a latent variables δ_i via f_3 for each data point ($i = 1, \ldots, n$) to construct a skewed tensor-t models, where $\delta_i/\nu \sim Ga(\frac{\nu}{2}, \frac{\nu}{2})$ of f_3. When we set $\boldsymbol{\delta} = (1, \ldots, 1)^\intercal \in \mathbb{R}^n$, and omit the parameter of degrees of freedom ν, we obtain the skewed tensor normal distribution. The corresponding joint posterior density is given by

$$p(\Theta|\mathcal{Y}) \propto L(\Theta|\mathcal{Y})\pi(\mathcal{B})\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\sigma})\pi(\nu)\pi(\boldsymbol{\rho}) \ , \tag{3.17}$$

where the $L(\Theta|\mathcal{Y})$ is the likelihood in (3.16), and $\pi(\cdot)$ represents independent marginal prior densities of $\boldsymbol{\lambda}, \boldsymbol{\sigma}$, ν, and $\boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is the vector of unknown parameters in $(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$. These priors and corresponding Bayesian computation based on the posterior in (3.17) will be discussed in the sections 3.5.3 - 3.5.5.

3.5.2 Incorporating Missing Responses

To easily accommodate missing responses within likelihood, we use the desirable property that our model is closed under marginalization over a specific mode of missing response (See, Corollary 1). For instance, in GAAD study, missing tooth implies that corresponding 6 sites and two biomarkers are missing. Define $\mathcal{W}_i \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the missing data indicator tensor with entries $w_{i,i_1i_2i_3} = 1$ if $y_{i,i_1i_2i_3}$ is missing, $w_{i,i_1i_2i_3} = 0$ if $y_{i,i_1i_2i_3}$ is observed. By abuse of notation, we denote $\mathcal{Y}_{i,c} = \{\mathcal{Y}_{\text{i,obs}}, \mathcal{Y}_{\text{i, mis}}\}$, where $\mathcal{Y}_{\text{i,obs}}$ is observed tensor response and $\mathcal{Y}_{\text{mis}}$ represents the missing part. Under either random (MAR) assumptions, we can ignore the missing mechanism (distribution of $\mathcal{W}_i$), and the likelihood contribution from $\mathcal{Y}_{\text{i, obs}}$ in (3.16) is proportional to $f(\mathcal{Y}_{i,\text{obs}}|\Theta) = \int f(\mathcal{Y}_{\text{i, obs}}, \mathcal{Y}_{\text{i, mis}}|\Theta) d\mathcal{Y}_{\text{i, mis}}$. Corollary 1 is a direct consequence of Theorem 3, and it guarantees that we can easily obtain the likelihood contribution $f(\mathcal{Y}_{\text{i, obs}}|\Theta)$ of any observed tensor response $\mathcal{Y}_{\text{i, obs}}$ without integrating out the $\mathcal{Y}_{\text{i, mis}}$ numerically.

Corollary 1. *Partition $\mathcal{Y}$ into $\mathcal{Y}^{(\ell)} \in \mathbb{R}^{d_{1,\ell} \times d_2 \cdots \times d_K}$ for $\ell = 1, 2$, $d_1 = d_{1,1} + d_{1,2}$, and $\boldsymbol{R}_1 = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{pmatrix}$ with $\boldsymbol{R}_{11}$ of dimension $(d_{1,1} \times d_{1,1})$. If $\mathcal{Y} \sim TSN(\boldsymbol{0}, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{K-1}, \boldsymbol{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$ as defined in (3.13), then $\mathcal{Y}^{(1)} \sim TSN(\boldsymbol{0}, \boldsymbol{R}_{11}, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_K; \boldsymbol{\sigma}, \boldsymbol{\lambda})$.*

A straightforward variation of Corollary 1 is valid even if we partition the tensor response along any specific mode. This property is useful especially for GAAD study, where a missing tooth (mode-1) of a patient has corresponding all missing sites and biomarkers. For GAAD study, where $\mathcal{Y}_i \in \mathbb{R}^{28 \times 6 \times 2}$, Corollary 1 gives us the observed response distribution $\mathcal{Y}_{i,\text{obs}} \sim STN(\mathcal{B}^{(1)} \bar{\times}_{(K+1)} \mathbf{x}_i; \mathbf{R}_{11}, \mathbf{R}_2, \mathbf{R}_3, \boldsymbol{\sigma}, \boldsymbol{\lambda})$ used for likelihood contribution in (3.16) via corresponding partition of $(\mathcal{B}, \mathbf{R}_1)$.

3.5.3 Prior Specification

For each subject with a K-way tensor response $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ and vector of covariates (including intercept) $\mathbf{x}_i \in \mathbb{R}^p$, the priors of all the unknown parameters Θ in our model (3.12)-(3.13) would be specified as follows.

(i) The prior $\pi(\mathcal{B})$ for tensor coefficient $\mathcal{B}$ is tensor normal assuming every component $\beta_{i_i,\ldots,i_K,j}$ of $\mathcal{B}$ is potentially different with each other. For instance, $\pi(\mathcal{B}) \sim TN(0; \mathbf{C}_1, \ldots, \mathbf{C}_{K+1})$ with prior mean zero tensor with $(K+1)$ known covariance matrices $\mathbf{C}_1, \ldots, \mathbf{C}_{K+1}$.

(ii) In (3.13), $\mathbf{\Lambda}_K = \text{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_K})$ and $d_K = 2$ in GAAD study. Then, λ_1 and λ_2 represent skewness parameters for PPD and CAL, respectively. The prior $\pi(\lambda_{i_K})$, for $i_K = 1, \ldots, d_K$ is univariate normal with mean zero with pre-specified variance. Using Theorem 3, each skewness parameter in $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ for GAAD study would be specified marginally. For example, $\boldsymbol{\lambda} \sim N\left(\begin{pmatrix} \mu_{\lambda_1} \\ \mu_{\lambda_2} \end{pmatrix}, \begin{pmatrix} c_{\lambda_1} & 0 \\ 0 & c_{\lambda_2} \end{pmatrix}\right)$, where $\mu_{\lambda_1} \neq \mu_{\lambda_2}$, and $c_{\lambda_1} \neq c_{\lambda_2}$.

(iii) In GAAD study, a prior for scale parameters is denoted by $\pi(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$. We independently specify each parameter as inverse-gamma distribution, $IG(g_1, g_2)$ with the shape parameter $g_1 > 0$, and scale parameter $g_2 > 0$. Each parameter represents the variability of PPD and CAL, respectively.

(iv) As we discussed in section 3.4.2, degrees of freedom should be sampled closing to two. So, we adapt the reference prior introduced by [53] in our setting, $\pi(\nu) \propto \frac{1}{\sigma_1\sigma_2}\sqrt{\psi'(\frac{\nu}{2}) - \psi'(\frac{\nu+1}{2}) - \frac{2(\nu+3)}{\nu(\nu+1)^2}}$, where $\psi(\nu) = \frac{d}{d\nu}\{\log\Gamma(\nu)\}$ and $\psi'(\nu) = \frac{d}{d\nu}\{\psi(\nu)\}$ are the digamma and trigamma functions, respectively. The reference prior performances very well for when the value of ν is small [25]. The details for the proposal distribution can be found in Appendices.

(v) The correlation matrices $\mathbf{R}_1, \ldots, \mathbf{R}_K$ are parametrized with corresponding unknown parameters $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$ in (3.17). For example, when $\mathbf{R}_k$ is a equicorrelation matrix with unknown off-diagonal element ρ_k, we can use independent Uniform$(0, 1)$ as prior $\pi(\rho_k)$ for $k = 1, \cdots, K$.

We implement MCMC sampling following the conditional posterior distributions in section 3.5.5. We sample $\mathcal{B}$, λ_{i_K}, $\sigma^2_{i_K}$, ρ_k, ν and $z_{2i,i_1\ldots i_K}$ straightforwardly within Gibbs sampler for $k = 1, \ldots, K$ and $i_K = 1, \ldots, d_K$. The kernel of posterior distribution of $\boldsymbol{\rho}$ and ν are not a standard closed form, thus we update $\boldsymbol{\rho}$ and ν using Metropolis-Hastings (MH) algorithm ([11]).

3.5.4 Tensor Spike-and-Slab Lasso Prior

Now, we introduce a practical shrinkage prior to evaluate overall covariate effect and detect elements of high signals different from other (estimable) sparse coefficients. This set-up is particularly useful for GAAD study in that a patient with diabetes are susceptible to bacterial infections in the buccal space [40]. The classical spike-and-slab prior [29] includes point-mass mixture, which may lead slow convergence of MCMC algorithm [43]. To tackle this issue, [47] proposed the spike-and-slab LASSO (SSL) prior with continuous relaxation of spike-and-slab prior. We extend SSL prior in the tensor coefficients accompanying efficient computation and refer it as tensor SSL (TSSL)

prior with the following parametrization.

$$\begin{aligned} \beta_{i_1 i_2 i_3 j} &= \eta_{i_3 j} + \alpha_{i_1 i_2 i_3 j} \text{ with } \eta_{i_3 j} \sim N(0, \tau^2) \text{ for } \tau > 0 \\ \pi(\alpha_{i_1 i_2 i_3 j} | \gamma_0, \gamma, \omega_{i_1 i_2 i_3 j}) &\overset{iid}{\sim} (1 - \omega_{i_1 i_2 i_3 j}) \psi(\alpha_{i_1 i_2 i_3 j} | \gamma_0) + \omega_{i_1 i_2 i_3 j} \psi(\alpha_{i_1 i_2 i_3 j} | \gamma), \\ (\omega_{i_1 i_2 i_3 j} | \phi) &\sim \text{Bernoulli}(\phi) \end{aligned} \tag{3.18}$$

where $\psi(\alpha_{i_1 i_2 i_3 j}|\gamma) = (\gamma/2)\exp(-\gamma\alpha_{i_1 i_2 i_3 j})$ is the Laplace distribution with mean zero and variance $2/\gamma^2$. When $\gamma \ll \gamma_0$, TSSL prior (3.18) closely retrieve to tensor spike-and-slab prior. The large value of spike penalty parameter produces highly sparse elements in $\alpha_{i_1 i_2 i_3}$. When the prior (3.18) chooses S non-zero elements of $\alpha_{i_1 i_2 i_3}$, then $\beta_{i_1 i_2 i_3 j}$ requires to estimate $(d_3 p + S)$ parameters in (3.18). We reduce $d_3 p(d_1 d_2 - 1) - S$ parameters compared to Tensor Normal prior. Also, other tensor regression models [23, 50] using Rank R PARAFAC decomposition [51] require $R(d_1 + d_2 + d_3 + p)$ parameters for tensor coefficient. The practical advantages of our model with TSSL prior will be demonstrated in GAAD data analysis at section 3.6.

3.5.5 Posterior Computation

We derive posterior distributions with Tensor Normal prior for regression coefficients for Bayesian skewed tensor-t (BSTT). Details are given in Appendices. Note that Bayesian tensor-t (BTT) is a special case of BSTT.

3.6 GAAD Data Analysis

The GAAD study was conducted by the Center for Oral Health Research (COHR) at the Medical University of South Carolina (MUSC). The study aims to investigate Type-2 diabetes of Gullah-speaking African-Americans adversely impacts PD status. Dental hygienists measured two biomarkers called PPD/CAL representing current status/progression of PD [14]. Biomarkers were measured in mm for each of the 6 sites per tooth (disto-buccal, mid-buccal, mesio-buccal, disto-lingual, mid-lingual and mesio-lingual), for a maximum of 28 teeth per participants (excluding the 4 third molars).

The main goal of data analysis is to evalutate overall effects of 5 covariates on both biomarkers: age (in years), gender (1=Female, 0=Male), Body Mass Index (obese if BMI $\geq$ 30, Not obese, otherwise), smoking status (1 = smoker, 0 = never), and HbA1c [12] level (Diabetes (coded as 1) if HbA1c $\geq$ 6.5, Not diabetes (coded as 0), otherwise) on PD status. We use $n = 290$ subjects with at

least one tooth available and complete covariate information. Thus, for GAAD study participant i, the regression function for each element of the tensor response is $Y_{i,i_1,i_2,i_3} = \sum_{j=1}^{p} \beta_{i_1,i_2,i_3,j} x_{ij} + \mathcal{E}_{i,i_1 i_2 i_3}$ from (3.12).

For data analysis, we consider 3 competing methods with TN and TSSL priors:

1. Bayesian tensor normal model (BTN): tensor normal density for $\mathcal{E}_i$ in (3.12) and $\mathbf{\Lambda} = \mathbf{0}$ in (3.13).
2. Bayesian tensor-t model (BTT): tensor-t density for $\mathcal{E}_i$ in (3.12) and $\mathbf{\Lambda} = \mathbf{0}$ in (3.13).
3. Bayesian skewed tensor-t model (BSTT): skewed tensor-t density for $\mathcal{E}_i$ in (3.12).

Three models also rapidly converge within first fifty samples as earlier works in the area of Bayesian tensor regression [22, 23, 49, 21]. Thus, we discard first 100 samples, and derive 1,000 samples with thinning of size 5 for all Bayesian models (BTN, BTT, BSTT). Parameters are mixed well over two independent chains.

3.6.1 Tensor Normal Prior

We employ Tensor Normal prior for regression coefficients (as specified in section 3.5.3) of 3 competing methods and illustrate the advantage of BSTT over other methods. Table 3.1 presents skewed tensor model (BSTT) is better than tensor models without accommodating skewness (BTN, BTT). Also, LPML [28] and WAIC [54] provide the BSTT (LPML $= -9.0710 \times 10^4$ and WAIC $= 24.4917 \times 10^4$) is the most appropriate model to analyze the GAAD data, followed by the BTT model (LPML $= -13.2734 \times 10^4$ and WAIC $= 55.8368 \times 10^4$). The LPML (WAIC) values for BTN model without adjusting heavy-tailedness are -14.3553×10^4 (11.3096×10^5), both considerably smaller (larger) than our new skewed tensor-t and tensor-t models. BTT and BSTT fittings show that the two biomarkers (PPD and CAL) reveal high correlations, with the estimated correlations (posterior median of ρ_3) 0.52, 0.45, respectively. Furthermore, BSTT captures high correlations among teeth (ρ_1) and among tooth-sites (ρ_2). The acceptance rate of sampling ν using MH algorithm within BTT/BSTT is 66.6%/70.2%. Also, BTT and BSTT detect significant non-Gaussuanity and heavy tails of GAAD data (see 95% CI of df parameter). We demonstrate that the estimated df parameter is near (but larger than) 2 as we discussed in section 3.4.2.

Furthermore, We compare 3 models with TN and TSSL priors using the statistic $D = d_1/|d_2|$ for $\beta_{i_1 i_2 i_3 j}$, where d_1 is the posterior median and d_2 is the distance of the posterior median from the nearest endpoint of the 95% CI of $\beta_{i_1 i_2 i_3 j}$. Here, $D > 1.5$ ($D < -1.5$) suggests strong posterior evidence of positive (negative) effect of each covariate on the biomarker at the site of the tooth.

Table 3.1: (2.5%, Median, 97.5%) of posterior estimates of parameters for three models

	BTN	BTT	BSTT
σ_1^2	(1.15, 1.17, 1.98)	(1.90, 2.30, 2.54)	(1.84, 2.24, 2.36)
σ_2^2	(1.61, 1.63, 2.66)	(2.17, 2.63, 2.90)	(2.06, 2.54, 2.79)
λ_1	-	-	(0.95, 1.15, 1.44)
λ_2	-	-	(1.31, 1.56, 1.75)
ρ_1	(0.09, 0.10, 0.20)	(0.08, 0.12, 0.19)	(0.50, 0.65, 0.79)
ρ_2	(0.20, 0.21, 0.39)	(0.17, 0.18, 0.25)	(0.38, 0.60, 0.77)
ρ_3	(0.23, 0.24, 0.60)	(0.32, 0.52, 0.55)	(0.30, 0.45, 0.63)
ν	-	(2.00003, 2.0013, 2.0067)	(2.0001, 2.0012, 2.0059)

Similarly, the range $0.8 \leq D \leq 1.5$ ($-1.5 \leq D \leq -0.8$) represents moderate positive (negative) effect of x_j on $\mathcal{Y}_{i_1 i_2 i_3}$, and $-0.8 < D < 0.8$ implies null evidence.

Figure 3.1 exhibits the heatmaps of D-statistics acquired from the BSTT model fit; each colored plate consists of 28 rows (teeth) and 6 columns (sites) on PPD (top row) and CAL (bottom row) measures. Each rectangular box demonstrates strength of evidence and the particular way of association of x_j on PPD or CAL ($i_3 = 1, 2$), for all 28×6 teeth-sites combinations. Figure 3.1 illustrates strong evidences of higher levels of PD on most of the teeth-site combinations are observed in elder, smokers, diabetic, males compared to younger, non-smokers, non-diabetic, and females, respectively. However, substantial plaid patterns are observed for all plates, so difficulties remain in interpreting overall covariate effects, feasibly due to the flatness of Tensor Normal prior on $\mathcal{B}$. Similar checkerboard patterns were also observed in the heatmaps using D-statistics from BTN and BTT model fits with the Tensor Normal prior on $\mathcal{B}$ (see Figures B.1 and B.2 in Appendices).

3.6.2 Tensor Spike-and-Slab Lasso Prior

To tackle difficulty of clinical interpretation with Tensor Normal prior, we further carry out Bayesian analysis with a sparsity inducing prior on the tooth-site specific effects ($\alpha_{i_1 i_2 i_3 j}$) to assess overall covariate effects ($\eta_{i_3 j}$) on two biomarkers ($i_3 = 1, 2$) in (3.18). Table 3.2 presents the posterior median, standard deviations (SD), and the 95% credible interval (CI)) of the overall effects $\eta_{i_3 j}$ (5 covariates and intercept terms on PPD and CAL) and skewness parameters obtained from fitting the BSTT method with TSSL prior for $\mathcal{B}$. Age has moderate/null posterior evidence on CAL/PPD. Posterior estimates of the effects of smoker, HbA1c on both PPD and CAL are positive, while BMI has not enough data evidence with 95% CIs including zero. Substantial posterior evidences of lower PPD and CAL are observed in women compared to male. The lower endpoints of CIs for both skewness parameters are larger than 0.9, revealing strong evidence of right-skewness for

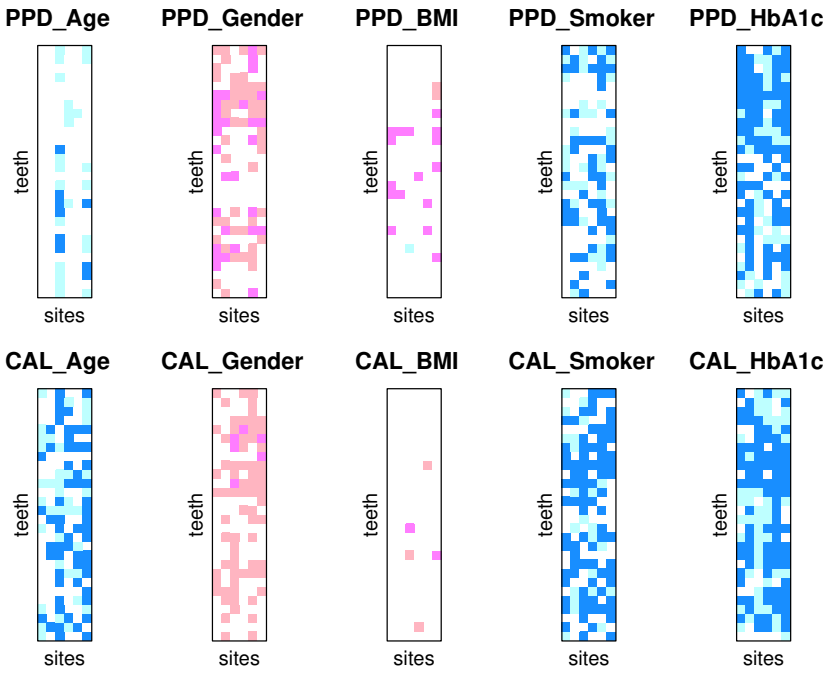

Figure 3.1: Fitting the BSTT Model with Tensor Normal prior for $\mathcal{B}$ to the GAAD data. Plotted are the D-statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

both responses. Compared to BSTT method with TSSL prior, the CIs for the bivariate regression model in [7] are wider for the skewness parameters and overall regression effects. The corresponding posterior summaries obtained from the BTN and BTT methods (with TSSL prior) can be found in Table 1 (Supplementary Materials). The 95% intervals for all parameters from BSTT methods with sparsity prior are narrower than BTN and BTT methods (with sparsity prior).

BSTT method with TSSL prior produces Figure 3.2 using D-statistics. It clearly shows the posterior evidence of covariate associations on both biomarkers for all individual tooth-site combinations compared to checkerboard pattern of Figure 3.1 from BSTT method with Tensor Normal prior. Strong posterior evidences of the positive associations of smoking and HbA1c on PPD and CAL are observed for (almost) all teeth-sites combinations, with a more prominent association with CAL. Age and smoking display similar strong evidence of positive associations on CAL for most

Table 3.2: Fitting the BSTT Model with TSSL prior for $\mathcal{B}$ to the GAAD data. Values in table are the posterior summaries of the overall covariate associations, and the skewness parameters, corresponding to the PPD (upper row), and CAL (lower row).

PPD	Median	SD	CI
Age	0.0016	0.0024	(-0.0029, 0.0066)
Gender	-0.1053	0.0631	(-0.2179, -0.0027)
BMI	0.0012	0.0559	(-0.1091, 0.1108)
Smoker	0.1615	0.0573	(0.0541, 0.2810)
HbA1c	0.2302	0.0591	(0.1183, 0.3446)
Skewness	0.8481	0.0485	(0.7609, 0.9438)
CAL	Median	SD	CI
Age	0.0108	0.0023	(0.0063, 0.0154)
Gender	-0.2151	0.0548	(-0.3232, -0.1066)
BMI	0.0326	0.0526	(-0.0688, 0.1339)
Smoker	0.3122	0.0515	(0.2111, 0.4123)
HbA1c	0.3039	0.0560	(0.1980, 0.4137)
Skewness	1.1005	0.0747	(0.9564, 1.2403)
ν	2.002	0.0021	(2.0001, 2.0077)

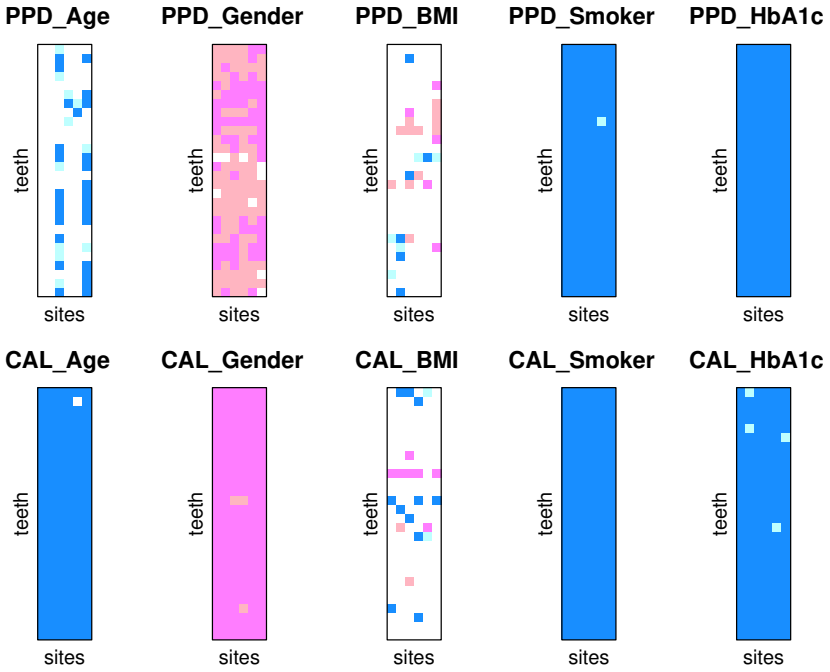

Figure 3.2: Fitting the BSTT Model with TSSL prior for $\mathcal{B}$ to the GAAD data. Plotted are the D-statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

teeth-sites combinations, however, age/smoking shows weak/moderate evidence of the positive associations on PPD. Gender shows strong negative association for both PPD and CAL, suggesting males experience worse PD status than females. On the contrary, Figures B.3 and B.4 in Appendices presenting the corresponding D-statistics heatmaps from the BTN and BTT fits (with TSSL prior) neglect to summarize covariate associations in a simple manner, particularly, age on PPD displays strong and null evidences together.

Table 3.3: Analysis of GAAD using BSTT with sparsity: Percentages of the posterior medians (Bayes point estimates) of $\alpha_{i_1 i_2 i_3 j} \neq 0$ in (3.18) for 6 tooth-sites combinations of interest. The highest percentage of tooth-sites over a covariate is highlighted in boldface.

PPD	disto-buccal	mid-buccal	mesio-buccal	disto-lingual	mid-lingual	mesio-lingual
Age	32.14%	39.29%	**78.57%**	32.14%	42.86%	75%
Gender	57.14%	50%	53.57%	35.71%	53.57%	**75%**
BMI	17.86%	21.43%	46.43%	21.43%	28.57%	**57.14%**
Smoker	42.86%	57.14%	**82.14%**	39.29%	57.14%	**82.14%**
HbA1c	**89.29%**	78.57%	78.57%	71.43%	50%	85.71%
CAL	disto-buccal	mid-buccal	mesio-buccal	disto-lingual	mid-lingual	mesio-lingual
Age	46.43%	39.29%	75%	39.29%	50%	**89.29%**
Gender	50%	46.43%	53.37%	50%	**60.71%**	**60.71%**
BMI	**35.71%**	25%	**35.71%**	28.57%	25%	14.29%
Smoker	71.43%	67.68%	85.71%	64.29%	71.43%	**92.86%**
HbA1c	**85.71%**	67.86%	71.43%	50%	67.86%	67.86%

Table 3.4: Analysis of GAAD using BSTT with sparsity: Percentages of the posterior medians (Bayes point estimates) of $\alpha_{i_1 i_2 i_3 j} \neq 0$ in (3.18) for 6 tooth-sites combinations of interest. The highest percentage of tooth-sites over a covariate is highlighted in boldface corresponding to fastest decaying teeth.

PPD	Molars	Pre-molars	Canines	Incisors
Age	52.08%	**58.33%**	33.33%	47.92%
Gender	33.33%	45.83%	**77.08%**	58.33%
BMI	**37.5%**	31.25%	33.33%	27.08%
Smoker	56.25%	**68.75%**	62.5%	54.17%
HbA1c	79.17%	**87.5%**	58.33%	68.75%
CAL	Molars	Pre-molars	Canines	Incisors
Age	**62.5%**	58.33%	54.17%	50%
Gender	43.75%	56.25%	**75%**	56.25%
BMI	**35.42%**	25%	29.17%	20.83%
Smoker	68.75%	**87.5%**	75%	68.75%
HbA1c	**77.08%**	75%	54.17%	50%

To detect fast decaying tooth-site and tooth-types, we set spike penalty parameter as 200 to produce highly sparse elements on $\alpha_{i_1 i_2 i_3 j}$ in (3.18) and slab variance parameter to be 1. We obtain the posterior median and 95% CIs of $\alpha_{i_1 i_2 i_3 j}$. Tables 3.3 and 3.4 present the percentages

of the posterior median of non-sparse elements ($\alpha_{i_1 i_2 i_3 j} \neq 0$), for the tooth-sites (combined across tooth-types) and tooth-types (packed over tooth-sites), respectively. While 'buccal' represents the sites towards the cheek, 'lingual' corresponds to the sites that are closest to the tongue. Also, 'mesial/distal' represents sites that are closest to/away from the midline of the buccal cavity (the gap sites), while 'mid' indicates the sites lying at the midpoint on either (buccal/lingual) sides. Higher glycemic status brings strong infection on the buccal area compared to the lingual sites, for both PPD and CAL responses, and existing finding [40] confirms our results. To identify susceptible tooth-types in Table 3.4, the 28 teeth (excluding the third molars) are classified as molars (#: 1-2, 13-16, 27-28), pre-molars (#: 3-4, 11-12, 17-18, 25-26), canines (#: 5, 10, 19, 24), and incisors (#: 6-9, 20-23). The effects of Age and HbA1c are the most prominent on the molars, followed by pre-molars, and other tooth types. Previous analyses of the GAAD data showed a relatively high proportion of diseased and missing molars [46], and the current finding confirms this fact. The interesting finding is that the effect of smoking is the highest on the pre-molars, followed by canines for both PPD and CAL responses.

3.7 Discussion

In this article, we develop a general elliptical class of skewed tensor distributions based on two findings; 1) any linear combinations of tensor elliptically distributed variables are tensor elliptical variables, 2) marginal density has the same form as the conditional density of skewed tensor elliptical distribution. The distribution of any subset of skewed tensor response is within the same skewed tensor distribution class of the original tensor. Thus, the density of subset of tensor response is a function of relevant skewness and scale parameters. The new class contains two special cases; a) Skewed Tensor Normal distribution (STN), b) Skewed Tensor-t distribution (STT). We also elicit useful properties of each distribution. We apply defined skewed distributions to tensor response regression analysis as demonstrated in GAAD data.

This paper is the first attempt to define a general class of skewed tensor distributions. There are various ways to define an elliptical class of skewed tensor distributions. Unlike transformation method we deployed, one may utilize conditioning method to define another class of skewed tensor distributions. This will be promising avenue for future research.

CHAPTER 4

FUTURE WORK

4.1 Possible Extension for Bayesian Skewed Tensor Normal Model

There exists a number of viable directions to extend our BSTN framework. For example, in the BSTN setup, missing values in the tensor response was integrated out, under the assumption of an ignorable MAR Bayesian framework. In the context of PD (and the GAAD dataset), tooth missingness is often conjectured to be informative/non-random [46], given that past incidence of PD is a leading cause of tooth failure. However, incorporating a missing-not-at-random [37] feature can complicate the current tensor setup due to varying dimensions imposed when subjects have different numbers of missingness for the teeth and sites.

Finally, our STN framework can also be extended to the general scenario as

$$\mathcal{E}_i = |\mathcal{Z}_{2i}| \times_1 \mathbf{\Lambda}_1 \cdots \times_K \mathbf{\Lambda}_K + \mathcal{Z}_{1i}$$

where, $\mathbf{\Lambda}_j = \text{diag}(\lambda_{j1}, \ldots, \lambda_{jd_j})$ for $j = 1, \ldots, K$. Unlike (2.2), here, each element $e_{i,i_1 \ldots i_K}$ of tensor $\mathcal{E}_i$ has its skewing parameter $\prod_{k=1}^{K} \lambda_{ki_k}$. For the GAAD study, this model allows the skewness levels to be different over teeth, sites, and two biomarkers, however, at the expense of the burden of estimating $\sum_{k=1}^{K} d_k$ number of skewness parameters. These interesting avenues will be considered elsewhere.

4.2 Bayesian Regression Analysis of Mixed-Type Matrix-variate Responses

In our current setup, we consider the tensor components to be continuous. However, in real data situations, they may constitute a variety of data types, such as discrete (such as bleeding on probing in PD), ordinal, etc. A latent variable formulation maybe worthwhile here for developing a unified analytical framework. Thus, we consider a multi-type $(d_1 \times d_2)$-dimensional matrix-variate responses $\mathbf{Y}_{i,\ell}$ for $i = 1, \ldots, n$, $\ell = 1, 2, 3$, where $\mathbf{Y}_{i,1}$, $\mathbf{Y}_{i,2}$ are continuous variables and $\mathbf{Y}_{i,3}$ is

binary variable. We model binary response via probit regression. For a p-dimensional vector of covariates $\mathbf{x}_i$, the regression model for multi-type responses is

$$\begin{cases} \mathbf{Y}_{i,\ell} & = \mathcal{B}_\ell \bar{\times}_3 \mathbf{x}_i + \eta_\ell \mathbf{W}_i + \lambda_\ell |\mathbf{Z}_{i,\ell}| + \mathbf{E}_{i,\ell} \text{ for } \ell = 1, 2 \\ \mathbf{Y}_{i,3} & = I(\mathbf{Y}^*_{i,3} > 0),\ \mathbf{Y}^*_{i,3} = \mathcal{B}_3 \bar{\times}_3 \mathbf{x}_i + \eta_3 \mathbf{W}_i + \mathbf{E}_{i,3}, \end{cases} \tag{4.1}$$

where $I(\cdot)$ is the binary indicator function, and $\mathbf{Y}^*_{i,3}$ is a matrix-variate Gaussian latent variable. For identifiability of the model for binary response, we assume $\mathbf{E}_{i,3} \sim MN(\mathbf{0}, \mathbf{I}_{d_1}, \mathbf{I}_{d_2})$. We can express binary response element-wisely, $\Phi^{-1}[Pr\{y_{ijk,3} = 1\}] = b_{jk,3}\mathbf{x}_i + \eta_3 w_{ijk} + e_{ijk,3}$, where $\Phi(\cdot)$ denotes the cumulative density function (cdf) for a standard normal distribution. $\mathcal{B}_\ell \in \mathbb{R}^{d_1 \times d_2 \times p}$ is a three-way unknown tensor of regression coefficients, $\bar{\times}_3$ is the third-mode vector product, and η_ℓ controls unknown matrix-variate latent variable $\mathbf{W}_i$, which has standard matrix normal distribution. A latent variable $\mathbf{W}_i$ plays crucial role of joint modeling due to common dependence of all responses. The parameter η_ℓ controls dependence among $\mathbf{Y}_{i,1}$, $\mathbf{Y}_{i,2}$, and $\mathbf{Y}_{i,3}$. Skewness parameters are denoted by λ_ℓ, and skewing shock matrix is $|\mathbf{Z}_{2i,\ell}|$, where $\mathbf{Z}_{2i,\ell} \sim MN(\mathbf{0}, \mathbf{I}_{d_1}, \mathbf{I}_{d_2})$ for $\ell = 1, 2$. The matrix-variate errors for continuous variables are denoted by $\mathbf{E}_{i,\ell} \sim MN(\mathbf{0}, \sigma^2_\ell \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ for $\ell = 1, 2$. They are assumed to be independent of latent variable $\mathbf{W}_i$.

APPENDIX A

APPENDIX OF CHAPTER 2

A.1 Lemmas

Lemma 6. *Under the vectorized skewed matrix-variate model with Kronecker covariance structure* $\boldsymbol{\Sigma} = \sigma^2\{\boldsymbol{R}_2 \otimes \boldsymbol{R}_1\} > 0$*, where* $\boldsymbol{R}_1$*, and* $\boldsymbol{R}_2$ *are positive definite if and only if* $\rho_1 \in (\frac{-1}{t-1}, 1)$*, and* $\rho_2 \in (\frac{-1}{s-1}, 1)$*, then equicorrelation matrices (*$\boldsymbol{R}_1$ *and* $\boldsymbol{R}_2$*) have closed forms as follows.*

$T \times T$ *matrix* $\boldsymbol{R}_1^{-1}$ *has diagonal elements* a *and off-diagonal elements* a'*, where* $a = \frac{1}{1-\rho_1}\big(1 - \frac{\rho_1}{1+(T-1)\rho_1}\big)$*,* $a' = -\frac{\rho_1}{(1-\rho_1)(1+(T-1)\rho_1)}$*.*

Proof of Lemma 6 Following the basic property of Kronecker product,

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\{\mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\}, \text{where } \mathbf{R}_1 \text{ has three types of structure.}$$

$\mathbf{R}_1^{-1}$ is defined by $\mathbf{R}_1 = (1-\rho_1)\mathbf{I}_t + \rho_1\mathbf{J}_{t\times t}$, where ρ_1 is correlation, $\mathbf{I}_t$ denotes the t-dimensional identity matrix, and $\mathbf{J}_{t\times t}$ is a $t \times t$ matrix of ones. The inverse of $\mathbf{R}_1$ is represented by

$$\mathbf{R}_1^{-1} = \frac{1}{1-\rho_1}\left(\mathbf{I}_t - \frac{\rho_1}{1+(t-1)\rho_1}\mathbf{J}_{t\times t}\right),$$

where $\det(\mathbf{R}_1^{-1}) = 1/\{(1-\rho_1)^{t-1}(1+(t-1)\rho_1)\}$ ([17]). In GAAD study, t represents number of teeth. Thus, $\mathbf{R}_1^{-1}$, denoted by

$$\begin{pmatrix} a & a' & \cdots & \cdots & a' \\ a' & a & a' & \cdots & a' \\ a' & a' & a & a' & \cdots \\ \vdots & \vdots & a' & \ddots & \\ a' & \cdots & \cdots & a' & a \end{pmatrix},$$

where $a = \frac{1}{1-\rho_1}\big(1 - \frac{\rho_1}{1+(t-1)\rho_1}\big), a' = -\frac{\rho_1}{(1-\rho_1)(1+(t-1)\rho_1)}$. □

Lemma 7. *Let A be a $d \times p$ matrix and $X \sim N_p(\mu, \Sigma)$, then $AX \sim N_d(A\mu, A\Sigma A^\intercal)$. Suppose $\mathcal{Y} \in \mathbb{R}^{T\times S\times J}$, and $\mathcal{Y} = \mathcal{X} \times_1 A_1 \times_2 A_2 \times_3 A_3$, where $\mathcal{X} \sim TN_{T,S,J}(\boldsymbol{0}, \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3)$. Let $A_1 \in \mathbb{R}^{T\times T}$, $A_2 \in \mathbb{R}^{S\times S}$, and $A_3 \in \mathbb{R}^{J\times J}$ be invertible matrices. Then, $\mathcal{Y} \sim TN_{T,S,J}(\boldsymbol{0}; \Sigma_1, \Sigma_2, \Sigma_3)$, where $\Sigma_1 = A_1 \boldsymbol{V}_1 A_1^\intercal$, $\Sigma_2 = A_2 \boldsymbol{V}_2 A_2^\intercal$, and $\Sigma_3 = A_3 \boldsymbol{V}_3 A_3^\intercal$.*

Proof of Lemma 7 Since $E[\mathcal{Y}] = E[\mathcal{X} \times_1 A_1 \times_2 A_2 \times_3 A_3] = E[\mathcal{X}] \times_1 A_1 \times_2 A_2 \times_3 A_3$ which follows that $E[\mathcal{X}] = \mathbf{0}$ due to the assumption that $\mathcal{X} \sim TN(\mathbf{0}, \mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)$.
We need to show that

$$\mathbf{Y}_{(1)} \sim MN_{T,SJ}(\mathbf{0}_{(1)}; \Sigma_1, \Sigma_2 \otimes \Sigma_3)$$
$$\mathbf{Y}_{(2)} \sim MN_{S,TJ}(\mathbf{0}_{(2)}; \Sigma_2, \Sigma_1 \otimes \Sigma_3)$$
$$\mathbf{Y}_{(3)} \sim MN_{J,TS}(\mathbf{0}_{(3)}; \Sigma_3, \Sigma_1 \otimes \Sigma_2)$$

Among above statement, we only need to show the first equation holds and the other two cases are dealt with the same way. By $\mathcal{Y} = \mathcal{X} \times_1 A_1 \times_2 A_2 \times_3 A_3$, we have

$$\mathbf{Y}_{(1)} = A_1^\intercal \mathcal{X}_{(1)} (A_2 \otimes A_3)$$

When $\mathbf{C} \in \mathbb{R}^{T\times S}$ and $\Sigma_1 = A_1 \mathbf{V}_1 A_1^\intercal$ and $\Sigma_2 = A_2 \mathbf{V}_2 A_2^\intercal$. Then $\mathbf{C} \sim MN_{T,S}(\mathbf{0}; \Sigma_1, \Sigma_2)$ if and only if there exists a $\mathbf{Z} \sim MN_{T,S}(\mathbf{0}, \mathbf{I}_T, \mathbf{I}_S)$ such that $\mathbf{C} = A_1 \mathbf{Z} A_2^\intercal$. Then, $\mathbf{Y}_{(1)} \sim MN_{T,SJ}(\mathbf{0}_{(1)}; \Sigma_1, \Sigma_2 \otimes \Sigma_3)$ since $\Sigma_1 = \mathbf{A}\mathbf{V}_1\mathbf{A}^\intercal$ and all the rows of $\mathbf{Y}_{(1)}$ denoted by $\mathbf{Y}_{(1)\cdot}$ follows $N_{SJ}(\mathbf{0}_{(1)\cdot}; \Sigma_2 \otimes \Sigma_3)$ by the property of Kronecker product, $(A_2 \otimes A_3)^\intercal (A_2 \otimes A_3) = (A_2^\intercal A_2) \otimes (A_3^\intercal A_3) = \Sigma_2 \otimes \Sigma_3$. Similarly we can show that

$$\mathbf{Y}_{(2)} \sim MN_{S,TJ}(\mathbf{0}_{(2)}; \Sigma_2, \Sigma_1 \otimes \Sigma_3) \quad \mathbf{Y}_{(3)} \sim MN_{J,TS}(\mathbf{0}_{(3)}; \Sigma_3, \Sigma_1 \otimes \Sigma_2)$$

which concludes the proof. □

A.2 MCMC Details for Multivariate Skewed Response regression

We consider the following multivariate (T-variate) skewed response regression model,

$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i + \lambda |\mathbf{z}_{2i}| + \mathbf{z}_{1i}, \tag{A.1}$$

Let $\mathbf{w}_i = |\mathbf{z}_{2i}|$. Assume $\mathbf{z}_{1i} \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \sigma^2 \mathbf{R}_\rho$. The likelihood of the model (A.1) follows

$$L(\mathbf{B},\rho,\sigma,\lambda,\mathbf{w}_i|\mathbf{y}_i) \propto \det(\mathbf{\Sigma})^{-\frac{n}{2}}\exp\Big[-\frac{1}{2}\sum_{i=1}^{n}\big(\mathbf{y}_i-\mathbf{B}\mathbf{x}_i-\lambda\mathbf{w}_i\big)^{\intercal}\mathbf{\Sigma}^{-1}\big(\mathbf{y}_i-\mathbf{B}\mathbf{x}_i-\lambda\mathbf{w}_i\big)\Big]$$

$$\propto \det(\mathbf{\Sigma})^{-\frac{n}{2}}\times$$

$$\exp\Big[\sum_{i=1}^{n}\Big\{-\frac{1}{2}\mathbf{y}_i^{\intercal}\mathbf{\Sigma}^{-1}\mathbf{y}_i+\mathbf{y}_i^{\intercal}\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{x}_i+\mathbf{y}_i^{\intercal}\mathbf{\Sigma}^{-1}\lambda\mathbf{w}_i-\mathbf{w}_i^{\intercal}\lambda\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{x}_i-\frac{1}{2}\mathbf{x}_i^{\intercal}\mathbf{B}^{\intercal}\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{x}_i-\frac{1}{2}\mathbf{w}_i^{\intercal}\lambda\mathbf{\Sigma}^{-1}\lambda\mathbf{w}_i\Big\}\Big]$$

, where $\det(\cdot)$ represents the determinant.

Latent variable, $\mathbf{w}_i$ is given by

$$f(\mathbf{w}_i)\sim TrN_T(\mathbf{0},\sigma^2\mathbf{I}_T)\propto\frac{1}{\sigma^{2T}}\exp\Big[-\frac{1}{2\sigma^2}\mathbf{w}_i^{\intercal}\mathbf{w}_i\Big]\mathbf{I}_{(0,\infty)}(\mathbf{w}_i),$$

where "TrN_T" represents truncated T-variate normal distribution and σ^{2T} is the determinant of $\sigma^2\mathbf{I}_T$.

The prior specifications for the $\mathbf{B},\rho,\sigma^2,\lambda$ follow

$$\pi_1(\mathbf{B})\sim \mathrm{MN}_{T,p}(\mathbf{0},\mathrm{c}\mathbf{I}_T,\mathrm{c}\mathbf{I}_p)\propto\exp\Big[-\frac{1}{2}\mathrm{tr}\Big\{\frac{1}{\mathrm{c}^2}\mathbf{B}^{\intercal}\mathbf{B}\Big\}\Big]$$

$$\pi_2(\lambda)\sim \mathrm{N}(0,b^2)$$

$$\pi_3(\rho)\sim \mathrm{Unif}(0,1)$$

$$\pi_4\Big(\frac{1}{\sigma^2}\Big)\sim Ga(g_1,g_2),\ g_1,\ g_2>0$$

A.2.1 Conditional Posterior Distribution for Skewed Multivariate Response Case

1. The conditional posterior distribution for $\boldsymbol{\beta}=\mathrm{vec}(\mathbf{B})$ satisfies

$p(\boldsymbol{\beta}|-)\propto N_{Tp}\Big(\mathrm{A}^{-1}\Big\{\mathrm{vec}(S_{xy})-\mathrm{vec}(S_{xw})\Big\},\mathrm{A}^{-1}\Big)$,

where $S_{xy}=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i^{\intercal}\mathbf{\Sigma}^{-1},\ S_{xw}=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{w}_i^{\intercal}\lambda\mathbf{\Sigma}^{-1}, S_{xx}=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\intercal}$, $\mathrm{A}=\Big(S_{xx}\otimes\mathbf{\Sigma}^{-1}+\frac{1}{\mathrm{c}^2}\mathbf{I}_{Tp}\Big)$.

Then, we derive $\mathbf{B}$ by reshaping $\boldsymbol{\beta}$ into $T\times p$ matrix.

2. The conditional posterior distribution for w_{it} satisfies

$p(\mathrm{w}_{it}|-)\propto TrN\Big(\frac{E_t}{D_{tt}},\frac{1}{D_{tt}}\Big)I(w_{it}>0)$, where "TrN" represents truncated (univariate) normal distribution.

$D_{tt}=\frac{1}{\sigma^2}\big(\lambda^2\mathbf{R}_{\rho}^{-1}+\mathbf{I}_T\big)_{tt}$ and $E_t=\big\{\lambda\mathbf{\Sigma}^{-1}\big(\mathbf{y}_i-\mathbf{B}\mathbf{x}_i\big)\big\}_{it}-\sum_{t'\neq t}\big(\lambda^2\mathbf{\Sigma}^{-1}\big)_{tt'}\mathrm{w}_{it'}$.

3. The conditional posterior distribution for λ satisfies

$$p(\lambda|-) \propto N\left(\frac{B}{A}, \frac{b^2}{A}\right), \text{ where}$$
$$A = \left(b^2 \sum_{i=1}^{n} \mathbf{w}_i^\intercal \boldsymbol{\Sigma}^{-1} \mathbf{w}_i + 1\right) \text{ and } B = \sum_{i=1}^{n} \left(\{\mathbf{y}_i^\intercal - \mathbf{x}_i^\intercal \mathbf{B}^\intercal\} \boldsymbol{\Sigma}^{-1} \mathbf{w}_i b^2\right).$$

4. The conditional posterior distribution for ρ (assume true $\rho = 0.5$) with closed form for $\mathbf{R}_\rho$ satisfies

$$p(\rho| -)$$
$$\propto \det(\mathbf{R}_\rho)^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i - \lambda \mathbf{w}_i\right)^\intercal \frac{1}{1-\rho} \left(\mathbf{I}_T - \frac{\rho}{1+(T-1)\rho} \mathbf{1}_{T\times T}\right) \left(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i - \lambda \mathbf{w}_i\right)\right]$$
$$\times \mathbf{I}_{(0,1)}(\rho)$$

, where $\mathbf{1}_{T\times T}$ is a $T \times T$ matrix of ones. We apply MH algorithm to update ρ with proposal densities as Beta(2, 2) ([35]) with following steps.

Set the initial values $\rho^{(1)}$. For each $s = 2, \ldots, S$,

a) sample $\rho^{new} \sim q(\cdot|\rho^{(s-1)})$.

b) set $\rho^{(s)} = \rho^{new}$ with the probability $\min\left\{1, \alpha_1(\rho_1, \rho_1^{new})\right\}$,

where $\alpha(\rho^{(s-1)}, \rho^{new}) = \dfrac{p(\rho^{new}|-) q(\rho^{(s-1)}|\rho^{new})}{p(\rho^{(s-1)}|-) q(\rho^{new}|\rho^{(s-1)})}$, otherwise set $\rho^{(s)} = \rho^{(s-1)}$

5. Here, we assume the variance across T-dimension denoted by σ^2, so there is not an identifiability problem. Using a gamma prior distribution ($\text{Ga}(g_1, g_2)$), where $g_1 = g_2 = 2$ for $\frac{1}{\sigma^2}$ satisfies following full conditional distribution with closed form of association structure (Lemma 1).

$$p\left(\frac{1}{\sigma^2}\middle| - \right) \propto Ga\left(g_1 + nT,\ \frac{1}{2}S + \frac{1}{2}W + g_2\right),$$
$$\text{where } S = \sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i - \lambda \mathbf{w}_i\right)^\intercal \mathbf{R}_\rho^{-1} \left(\mathbf{y}_i - \mathbf{B}\mathbf{x}_i - \lambda \mathbf{w}_i\right), \text{ and } W = \sum_{i=1}^{n} \mathbf{w}_i^\intercal \mathbf{w}_i.$$

A.3 Details for MCMC Implementation for Skewed Matrix-variate Response Case

We consider the following matrix-variate skewed response regression model,

$$\mathbf{Y}_i = \mathcal{B}\bar{\times}_3\mathbf{x}_i + \lambda|\mathbf{Z}_{2i}| + \mathbf{Z}_{1i}, \tag{A.2}$$

and the vectorized model of (A.2) is

$$\text{vec}(\mathbf{Y}_i) = \mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i + \lambda\text{vec}(|\mathbf{Z}_{2i}|) + \text{vec}(\mathbf{Z}_{1i}), \tag{A.3}$$

where $\text{vec}(\mathbf{Z}_{1i}) \sim N(\text{vec}(\mathbf{0}), \sigma^2\{\mathbf{R}_2 \otimes \mathbf{R}_1\})$.

The likelihood function of vectorized model (A.3) is given by

$$\begin{aligned}
&L(\mathcal{B}, \text{vec}(|\mathbf{Z}_{2i}|), \sigma^2, \mathbf{R}_1, \mathbf{R}_2, \lambda|\text{vec}(\mathbf{Y}_i)) \\
&\propto \det(\mathbf{\Sigma})^{-\frac{n}{2}}\exp\bigg(-\frac{1}{2}\sum_{i=1}^{n}\bigg[\text{vec}(\mathbf{Y}_i) - \Big\{\mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i + \text{vec}(\lambda|\mathbf{Z}_{2i}|)\Big\}\bigg]^{\intercal}\mathbf{\Sigma}^{-1} \\
&\bigg[\text{vec}(\mathbf{Y}_i) - \Big\{\mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i + \text{vec}(\lambda|\mathbf{Z}_{2i}|)\Big\}\bigg]\bigg),
\end{aligned}$$

where $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\{\mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\}$. To derive full conditional distribution of parameters of interest, we multiply likelihood and prior distribution of each parameter and latent variable. For ease of notation, we define $\boldsymbol{\beta} = \text{vec}(\mathcal{B}), \mathbf{w}_i = \text{vec}(|\mathbf{Z}_{2i}|)$, and $\mathbf{y}_i = \text{vec}(\mathbf{Y}_i)$ in the proof.

Latent variable, $\mathbf{w}_i$ is given by

$$f(\mathbf{w}_i) \sim TrN_{TS}(\text{vec}(\mathbf{0}), \sigma^2\mathbf{I}_{TS}) \propto \frac{1}{\sigma^{2T}}\exp\bigg[-\frac{1}{2\sigma^2}\mathbf{w}_i^{\intercal}\mathbf{w}_i\bigg]\mathbf{I}_{(0,\infty)}(\mathbf{w}_i),$$

where "TrN_{TS}" represents truncated TS-variate normal distribution and σ^{2TS} is the determinant of $\sigma^2\mathbf{I}_{TS}$.

The prior specifications for the $\boldsymbol{\beta}, \rho_1, \rho_2, \sigma^2, \lambda$ follow

$$\pi_1(\boldsymbol{\beta}) \sim \text{N}_{TSp}(\text{vec}(\mathbf{0}), \text{c}\mathbf{I}_T, \text{c}\mathbf{I}_S, \text{c}\mathbf{I}_p) \propto \exp\bigg[-\frac{1}{2}\text{tr}\bigg\{\frac{1}{\text{c}^3}\boldsymbol{\beta}^{\intercal}\boldsymbol{\beta}\bigg\}\bigg]$$

$$\pi_2(\lambda) \sim \text{N}(0, b^2)$$

$$\pi_{3,k}(\rho_k) \sim \text{Unif}(0,1), \text{ for } k = 1, 2$$

$$\pi_4\bigg(\frac{1}{\sigma^2}\bigg) \sim Ga(g_1, g_2), \ g_1, \ g_2 > 0$$

A.3.1 Conditional Posterior Distribution for Skewed Matrix-variate Response Case

1. The full conditional distribution for $\boldsymbol{\beta} = \text{vec}(\mathbf{B}_{(3)}^{\intercal})$ using vectorized form satisfies

$$p(\boldsymbol{\beta}|-) \propto N_{TSp}\bigg(\mathrm{A}^{-1}\bigg\{\text{vec}(S_{xy}) - \text{vec}(S_{xw})\bigg\}, \mathrm{A}^{-1}\bigg),$$

$$\text{where } S_{xy} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i^{\intercal}\boldsymbol{\Sigma}^{-1},\ S_{xw} = \sum_{i=1}^{n}\mathbf{x}_i\lambda\mathbf{w}_i^{\intercal}\boldsymbol{\Sigma}^{-1}, S_{xx} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\intercal},\ \mathrm{A} = \bigg(S_{xx}\otimes\boldsymbol{\Sigma}^{-1} + \frac{1}{\mathrm{c}^2}\mathbf{I}_{TSp}\bigg),$$

and $\boldsymbol{\Sigma} = \sigma^2\{\mathbf{R}_2\otimes\mathbf{R}_1\}$ and $\mathbf{B}_{(3)}$ is matricization along the third mode of $\mathcal{B}$.

2. The full conditional distribution for w_{its} is derived as follows:

$$p(\mathrm{w}_{its}|-) \propto TrN\bigg(\frac{E_{ts}}{D_{ts,ts}}, \frac{1}{D_{ts,ts}}\bigg)I(w_{its} > 0),\ \text{where}$$

$$D_{ts,ts} = \big\{\lambda^2\boldsymbol{\Sigma}^{-1} + \mathbf{I}_{TS}\big\}_{ts,ts} \text{ and } E_{ts} = \big\{\lambda\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i)\big\}_{its} - \sum_{ts\neq t's'}(\lambda^2\boldsymbol{\Sigma}^{-1})_{ts,t's'}\mathrm{w}_{it's'},$$

and subscript (ts, ts) represents ts-th diagonal element.

3. The conditional posterior distribution for λ satisfies

$$p(\lambda|-) \propto N\bigg(\frac{B}{A}, \frac{b^2}{A}\bigg),\ \text{where}$$

$$A = \bigg(b^2\sum_{i=1}^{n}\mathbf{w}_i^{\intercal}\boldsymbol{\Sigma}^{-1}\mathbf{w}_i + 1\bigg) \text{ and } B = \sum_{i=1}^{n}\bigg(\{\mathbf{y}_i^{\intercal} - \mathbf{x}_i^{\intercal}\mathbf{B}_{(3)}\}\boldsymbol{\Sigma}^{-1}\mathbf{w}_i b^2\bigg).$$

4. The full conditional distributions for ρ_1 and ρ_2 which prior distributions are Uniform(0,1) follow

$$f(\rho_1, \rho_2|-)$$

$$\propto \det(\mathbf{R}_2\otimes\sigma^2\mathbf{R}_1)^{-\frac{n}{2}}\exp\bigg[-\frac{1}{2}\sum_{i=1}^{n}\big(\mathbf{y}_i - \mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i - \lambda\mathbf{w}_i\big)^{\intercal}\frac{1}{\sigma^2}\{\mathbf{R}_2^{-1}\otimes\mathbf{R}_1^{-1}\}\big(\mathbf{y}_i - \mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i - \lambda\mathbf{w}_i\big)\bigg]$$

$$\times\mathbf{I}_{(0,1)}(\rho_1)\mathbf{I}_{(0,1)}(\rho_2)$$

Let $\boldsymbol{\rho} = (\rho_1, \rho_2)$, and $\boldsymbol{\rho}^* = (\rho_1^*, \rho_2^*)$. We apply MH algorithm to update ρ_1 and ρ_2 jointly with proposal densities as $\rho_1^*,\ \rho_2^* \sim Beta(2, 2)$.

Set the initial values $\rho_1^{(1)}$ and $\rho_2^{(1)}$. For each $s = 2, \ldots, S$,

a) sample $\rho_1^{new} \sim q(\cdot|\rho_1^{(s-1)})$ and $\rho_2^{new} \sim g(\cdot|\rho_2^{(s-1)})$.

b) set $\rho_1^{(s)} = \rho_1^{new}$ with the probability $\min\left\{1, \alpha_1\left(\rho_1^{(s-1)}, \rho_1^{new}\right)\right\}$,

where $\alpha_1(\rho_1^{(s-1)}, \rho_1^{new}) = \dfrac{p(\rho_1^{new}|-)q(\rho_1^{(s-1)}|\rho_1^{new})}{p(\rho_1^{(s-1)}|-)q(\rho_1^{new}|\rho_1^{(s-1)})}$,

otherwise set $\rho_1^{(s)} = \rho_1^{(s-1)}$.

c) set $\rho_2^{(s)} = \rho_2^{new}$ with the probability $\min\left\{1, \alpha_2\left(\rho_2^{(s-1)}, \rho_2^{new}\right)\right\}$,

where $\alpha_2(\rho_2^{(s-1)}, \rho_2^{new}) = \dfrac{p(\rho_2^{new}|-)q(\rho_2^{(s-1)}|\rho_2^{new})}{p(\rho_2^{(s-1)}|-)q(\rho_2^{new}|\rho_2^{(s-1)})}$,

otherwise set $\rho_2^{(s)} = \rho_2^{(s-1)}$.

5. We only need to derive full-conditional distribution for σ^2 which implies variance of a bio-marker.

$$p\left(\frac{1}{\sigma^2}\middle| - \right) \propto Ga\left(g_1 + nTS,\ \frac{1}{2}S^* + \frac{1}{2}W + g_2\right),$$

where $S^* = \sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i - \lambda\mathbf{w}_i\right)^{\intercal}\{\mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\}\left(\mathbf{y}_i - \mathbf{B}_{(3)}^{\intercal}\mathbf{x}_i - \lambda\mathbf{w}_i\right)$, and $W = \sum_{i=1}^{n} \mathbf{w}_i^{\intercal}\mathbf{w}_i$.

A.4 Details for MCMC Implementation for Skewed Tensor-variate Response Case

We derive full-conditional distributions for three way tensor response case. The likelihood function is given by

$$\begin{aligned}
&L(\mathcal{B}, \mathrm{vec}(|\mathbf{Z}_{2i}|), \mathbf{D}_\sigma, \mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \mathbf{\Lambda}|\mathrm{vec}(\mathcal{Y}_i)) \\
&\propto \det(\mathbf{\Sigma}^\star)^{-\frac{n}{2}} \exp\Bigg(-\frac{1}{2}\sum_{i=1}^{n}\left[\mathrm{vec}(\mathcal{Y}_{i(3)}^{\intercal}) - \left\{\mathbf{B}_{(4)}^{\intercal}\mathbf{x}_i + \mathrm{vec}(|\mathcal{Z}_{2i}| \times_3 \mathbf{\Lambda})\right\}\right]^{\intercal}\mathbf{\Sigma}^{\star -1} \\
&\times \left[\mathrm{vec}(\mathcal{Y}_{i(3)}^{\intercal}) - \left\{\mathbf{B}_{(4)}^{\intercal}\mathbf{x}_i + \mathrm{vec}(|\mathcal{Z}_{2i}| \times_3 \mathbf{\Lambda})\right\}\right]\Bigg),
\end{aligned}$$

where $\mathbf{B}_{(4)}^{\intercal}\mathbf{x}_i = \mathrm{vec}\{(\mathcal{B}\bar{\times}_4\mathbf{x}_i)_{(3)}^{\intercal}\}$, $\mathbf{\Sigma}^\star = \Sigma_3 \otimes \mathbf{R}_1 \otimes \mathbf{R}_2$, $\Sigma_3 = \mathbf{D}_\sigma\mathbf{R}_3\mathbf{D}_\sigma = \begin{pmatrix} \sigma_1^2 & \rho_3\sigma_1\sigma_2 \\ \rho_3\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, and $\mathbf{R}_1$, $\mathbf{R}_2$ have equicorrelation structures. For ease of notation, we define $\boldsymbol{\beta} = \mathrm{vec}(\mathcal{B})$, $\mathbf{y}_i = \mathrm{vec}(\mathcal{Y}_{i(3)}^{\intercal})$, $\mathbf{z}_{1i} = \mathrm{vec}(\mathcal{Z}_{1i(3)}^{\intercal})$, $\mathrm{vec}(|\mathcal{Z}_{2i}| \times_3 \mathbf{\Lambda}) = \mathrm{vec}(\mathbf{I}_{TS}|\mathbf{Z}_{2i(3)}|^{\intercal}\mathbf{\Lambda}) = (\mathbf{\Lambda} \otimes \mathbf{I}_{TS})\mathrm{vec}(|\mathbf{Z}_{2i(3)}|^{\intercal}) = (\mathbf{\Lambda} \otimes \mathbf{I}_{TS})\mathbf{w}_i$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2)$ in the proof.

$$\mathbf{y}_i = \mathbf{B}_{(4)}^{\intercal}\mathbf{x}_i + (\mathbf{\Lambda} \otimes \mathbf{I}_{TS})\mathbf{w}_i + \mathbf{z}_{1i}, \quad \text{for} \quad i = 1, \ldots, n, \tag{A.4}$$

The exponential term in the likelihood is expressed as follows:

$$\exp\Bigg(-\frac{1}{2}\sum_{i=1}^{n}\Big[\mathbf{y}_i^\intercal\boldsymbol{\Sigma}^{\star-1}\mathbf{y}_i - 2\mathbf{y}_i^\intercal\boldsymbol{\Sigma}^{\star-1}\mathbf{B}_{(4)}^\intercal\mathbf{x}_i - 2\mathbf{y}_i^\intercal\boldsymbol{\Sigma}^{\star-1}(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\mathbf{w}_i + 2\mathbf{w}_i^\intercal(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\boldsymbol{\Sigma}^{\star-1}\mathbf{B}_{(4)}^\intercal\mathbf{x}_i$$
$$+\mathbf{x}_i^\intercal\mathbf{B}_{(4)}\boldsymbol{\Sigma}^{\star-1}\mathbf{B}_{(4)}^\intercal\mathbf{x}_i + \mathbf{w}_i^\intercal(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\boldsymbol{\Sigma}^{\star-1}(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\mathbf{w}_i\Big]\Bigg)$$

The latent variable, $\mathbf{w}_i$ is specified as

$$f(\mathbf{w}_i) \sim TrN_{TSJ}(\mathbf{0}, \mathbf{D}_\sigma^2\otimes\mathbf{I}_{TS}) \propto \det(\mathbf{D}_\sigma^2\otimes\mathbf{I}_{TS})^{-\frac{1}{2}}\exp\Big\{ -\frac{1}{2}\mathbf{w}_i^\intercal(\mathbf{D}_\sigma^2\otimes\mathbf{I}_{TS})^{-1}\mathbf{w}_i\Big\}\mathbf{I}_{(0,\infty)}(\mathbf{w}_i).$$

The prior specifications for the $\boldsymbol{\beta}, \boldsymbol{\Lambda}, \rho_1, \rho_2, \rho_3, \sigma_1^2, \sigma_2^2$ follow

$$\pi_1(\boldsymbol{\beta}) \sim \mathrm{N}_{TSJp}(\mathrm{vec}(\mathbf{0}), \mathrm{c}\mathbf{I}_p\otimes\mathrm{c}\mathbf{I}_B\otimes\mathrm{c}\mathbf{I}_S\otimes\mathrm{c}\mathbf{I}_T) \propto \exp\Big[-\frac{1}{2\mathrm{c}^4}\{\boldsymbol{\beta}^\intercal\boldsymbol{\beta}\}\Big]$$
$$\text{where } \mathbf{D}_\sigma^2 = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$$
$$\pi_2(\boldsymbol{\Lambda}) \sim \mathrm{N}_2(\mathbf{1}, \mathbf{I}_2), \text{ where } \mathbf{1} = (1,1)^\intercal$$
$$\pi_{3,l}(\rho_\ell) \sim \mathrm{Unif}(0,1), \text{ for } \ell = 1,2,3$$
$$\pi_4\Big(\frac{1}{\sigma_b^2}\Big) \sim Ga(g_1, g_2), \text{where } g_1 = g_2 > 0,\ b = 1,2$$

A.4.1 Conditional Posterior Distribution for Skewed Tensor-variate Response case

1. The posterior full conditional distribution for $\boldsymbol{\beta} = \mathrm{vec}(\mathcal{B}) = \mathrm{vec}(\mathbf{B}_{(4)}^\intercal)$ from the vectorized model (A.4) is given by

$$p(\boldsymbol{\beta}|-) \propto N_{TSJp}\Big(\mathrm{A}^{-1}\Big\{\mathrm{vec}(S_{xy}) - \mathrm{vec}(S_{xw})\Big\}, \mathrm{A}^{-1}\Big), \text{ where}$$
$$S_{xy} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i^\intercal\boldsymbol{\Sigma}^{\star-1},\ S_{xw} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{w}_i^\intercal(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\boldsymbol{\Sigma}^{\star-1}, S_{xx} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\intercal,\ \mathrm{A} = \Big(S_{xx}\otimes\boldsymbol{\Sigma}^{\star-1} + \frac{1}{\mathrm{c}^4}\mathbf{I}_{TSJp}\Big),$$

and $\boldsymbol{\Sigma}^\star = \Sigma_3\otimes\mathbf{R}_2\otimes\mathbf{R}_1$ and $\mathbf{B}_{(4)}$ is matricization along the fourth mode of $\mathcal{B}$.

2. Based on (A.4), the posterior full-conditional distribution for w_{itsj} satisfies

$$p(w_{itsj}|-) \propto TrN\Big(\frac{E_{tsj}}{D_{tsj,tsj}}\frac{1}{D_{tsj,tsj}}\Big)I(w_{itsj} > 0), \text{ where}$$
$$D_{tsj,tsj} = \big\{(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\boldsymbol{\Sigma}^{\star-1}(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS}) + (\mathbf{D}_\sigma^2\otimes\mathbf{I}_{TS})^{-1}\big\}_{tsj,tsj},$$
$$E_{tsj} = \{(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\boldsymbol{\Sigma}^{\star-1}(\mathbf{y}_i - \mathbf{B}_{(4)}^\intercal\mathbf{x}_i)\}_{itsj}$$
$$-\sum_{tsj\neq t's'j'}\big\{(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS})\boldsymbol{\Sigma}^{\star-1}(\boldsymbol{\Lambda}\otimes\mathbf{I}_{TS}) + (\mathbf{D}_\sigma^2\otimes\mathbf{I}_{TS})^{-1}\big\}_{tsj,t's'j'}\mathrm{w}_{it's'j'},$$

and subscript (tsj, tsj) stands for tsj-th diagonal element of $TSJ \times TSJ$ matrix.

3. Define $\boldsymbol{\lambda} = \text{vec}(\boldsymbol{\Lambda})$, and λ_j represents skewness parameter for jth biomarker. The posterior full conditional distribution for λ_j derived based on (A.4) as follows:

$$
\begin{aligned}
&p(\lambda_j) \propto N\left(\frac{G_\delta}{H_{\delta,\delta}}, \frac{1}{H_{\delta,\delta}}\right) \\
&\text{Let } H_{\delta,\delta} = \sum_{i=1}^{n} \left[(|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J)^\intercal \boldsymbol{\Sigma}^{\star -1} (|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J) + \mathbf{I}_{J^2}\right]_{j^2, j^2}, \\
&G_\delta = \sum_{i=1}^{n} \Big\{ \left[(\mathbf{y}_i^\intercal - \mathbf{x}_i \mathbf{B}_{(4)}) \boldsymbol{\Sigma}^{\star -1} (|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J) + \mathbf{1}_0\right]_{j^2} \\
&- \sum_{j^2 \neq j'^2} \left[(|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J)^\intercal \boldsymbol{\Sigma}^{\star -1} (|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J) + \mathbf{I}_{J^2}\right]_{j^2, j'^2} \lambda_{j'} \Big\}, \\
&\text{and } (j^2, j^2) \text{ represents } j^2\text{-th diagonal elements of } J^2 \times J^2 \text{ matrix.}
\end{aligned}
$$

4. The full conditional distributions for ρ_1, ρ_2, and ρ_3 which prior distributions are Uniform(0,1) follow

$$
\begin{aligned}
&f(\rho_1, \rho_2, \rho_3 | -) \\
&\propto \det(\mathbf{D}_\sigma \mathbf{R}_3 \mathbf{D}_\sigma \otimes \mathbf{R}_2 \otimes \mathbf{R}_1)^{-\frac{n}{2}} \exp\Bigg[-\frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i - (\boldsymbol{\Lambda} \otimes \mathbf{I}_{TS}) \mathbf{w}_i\right)^\intercal \\
&\times \left(\{\mathbf{D}_\sigma^{-1} \mathbf{R}_3^{-1} \mathbf{D}_\sigma^{-1}\} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\right) \left(\mathbf{y}_i - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i - (\boldsymbol{\Lambda} \otimes \mathbf{I}_{TS}) \mathbf{w}_i\right) \Bigg] \mathbf{I}_{(0,1)}(\rho_1) \mathbf{I}_{(0,1)}(\rho_2) \mathbf{I}_{(0,1)}(\rho_3)
\end{aligned}
$$

Let $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$, and $\boldsymbol{\rho}^* = (\rho_1^*, \rho_2^*, \rho_3^*)$. We apply MH algorithm to update ρ_1, ρ_2, and ρ_3 jointly with proposal densities as ρ_1^*, ρ_2^* $\rho_3^* \sim Beta(2, 2)$.

Set the initial values $\rho_1^{(1)}$, $\rho_2^{(1)}$ and $\rho_3^{(1)}$. For each $s = 2, \ldots, S$,

a) sample $\rho_1^{new} \sim q(\cdot | \rho_1^{(s-1)})$, $\rho_2^{new} \sim g(\cdot | \rho_2^{(s-1)})$, and $\rho_3^{new} \sim g(\cdot | \rho_3^{(s-1)})$.

b) set $\rho_1^{(s)} = \rho_1^{new}$ with the probability $\min\left\{1, \alpha_1\left(\rho_1^{(s-1)}, \rho_1^{new}\right)\right\}$,

where $\alpha_1(\rho_1^{(s-1)}, \rho_1^{new}) = \dfrac{p(\rho_1^{new}|-) q(\rho_1^{(s-1)} | \rho_1^{new})}{p(\rho_1^{(s-1)}|-) q(\rho_1^{new} | \rho_1^{(s-1)})}$,

otherwise set $\rho_1^{(s)} = \rho_1^{(s-1)}$.

c) set $\rho_2^{(s)} = \rho_2^{new}$ with the probability $\min\left\{1, \alpha_2\left(\rho_2^{(s-1)}, \rho_2^{new}\right)\right\}$,

where $\alpha_2(\rho_2^{(s-1)}, \rho_2^{new}) = \dfrac{p(\rho_2^{new}|-)q(\rho_2^{(s-1)}|\rho_2^{new})}{p(\rho_2^{(s-1)}|-)q(\rho_2^{new}|\rho_2^{(s-1)})}$,

otherwise set $\rho_2^{(s)} = \rho_2^{(s-1)}$.

d) set $\rho_3^{(s)} = \rho_3^{new}$ with the probability $\min\left\{1, \alpha_3\left(\rho_3^{(s-1)}, \rho_3^{new}\right)\right\}$,

where $\alpha_3(\rho_3^{(s-1)}, \rho_3^{new}) = \dfrac{p(\rho_3^{new}|-)q(\rho_3^{(s-1)}|\rho_3^{new})}{p(\rho_3^{(s-1)}|-)q(\rho_3^{new}|\rho_3^{(s-1)})}$,

otherwise set $\rho_3^{(s)} = \rho_3^{(s-1)}$.

5-1. The full-conditional distribution for $1/\sigma_1^2$ based on (A.4) satisfies

$$
\begin{aligned}
&p\left(\frac{1}{\sigma_1^2}\middle| - \right) \propto Ga(nTS + g_1, \nu), \text{ where} \\
&\nu = \frac{1}{2}\sum_{i=1}^{n}\left\{\mathbf{S}^{\intercal}\left[\begin{pmatrix} \frac{1}{(1-\rho_3)(1+\rho_3)} & -\frac{\sigma_1\rho_3}{\sigma_2(1-\rho_3)(1+\rho_3)} \\ -\frac{\sigma_1\rho_3}{\sigma_2(1-\rho_3)(1+\rho_3)} & \frac{\sigma_1^2}{\sigma_2^2(1-\rho_3)(1+\rho_3)} \end{pmatrix} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\right]\mathbf{S}\right. \\
&\left. + \text{vec}(|\mathcal{Z}_{2i}|)^{\intercal}\text{diag}(1, \sigma_1^2/\sigma_2^2, \ldots, 1, \sigma_1^2/\sigma_2^2)\text{vec}(|\mathcal{Z}_{2i}|)\right\} + g_2, \\
&\mathbf{S} = \text{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^{\intercal}\mathbf{x}_i - (\mathbf{\Lambda} \otimes \mathbf{I}_{TS})\text{vec}(|\mathcal{Z}_{2i}|).
\end{aligned}
$$

5-2. The full-conditional distribution for $1/\sigma_2^2$ based on (A.4) satisfies

$$
\begin{aligned}
&p\left(\frac{1}{\sigma_2^2}\middle| - \right) \propto Ga(nTS + g_3, \nu_2), \text{ where} \\
&\nu_2 = \frac{1}{2}\sum_{i=1}^{n}\left\{\mathbf{S}^{\intercal}\left[\begin{pmatrix} \frac{\sigma_2^2}{\sigma_1^2(1-\rho_3)(1+\rho_3)} & -\frac{\sigma_2\rho_3}{\sigma_1(1-\rho_3)(1+\rho_3)} \\ -\frac{\sigma_2\rho_3}{\sigma_1(1-\rho_3)(1+\rho_3)} & \frac{1}{(1-\rho_3)(1+\rho_3)} \end{pmatrix} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\right]\mathbf{S}\right. \\
&\left. + \text{vec}(|\mathcal{Z}_{2i}|)^{\intercal}\text{diag}(\sigma_2^2/\sigma_1^2, 1, \ldots, \sigma_2^2/\sigma_1^2, 1)\text{vec}(|\mathcal{Z}_{2i}|)\right\} + g_4.
\end{aligned}
$$

A.5 Additional Simulation Study: Matrix-variate Response Case

We generate the matrix variate response $\mathbf{Y}_i = \mathcal{B}\bar{\times}_3\mathbf{x}_i + \lambda|\mathbf{Z}_{2i}| + \mathbf{Z}_{1i}$, where the true coefficient $\mathcal{B} = \{\mathbf{B}_0, \mathbf{B}_1, \mathbf{B}_2\}$, with the dimension of $\mathcal{B} \in \mathbb{R}^{\{20\times6\times3, 10\times3\times3\}}$. Under **Model 1** (Non-low rank coefficient), we set the true $\mathcal{B} = \{\mathbf{B}_0, \mathbf{B}_1, \mathbf{B}_2\}$, where $\mathbf{B}_0$ is a zero matrix and elements of $\mathbf{B}_1$ and $\mathbf{B}_2$ are sampled from standard normal distribution. Under **Model 2** (Non-low rank coefficients with opposite signs), $\mathbf{B}_0$ is again a zero matrix, while the components of $\mathbf{B}_1$ and $\mathbf{B}_2$ are sampled from $N(-1, 1)$ and $N(1, 1)$, respectively. The rest of the setup for both models remain the same, with the skewness parameter $\lambda = 2$ when simulated data is skewed (Scenario ii), otherwise $\lambda = 0$ (Scenario i). The covariate vector $\mathbf{x}_i = (1, x_{1i}, x_{2i})^\intercal \in \mathbb{R}^3$, where binary covariates $x_{1i}, x_{2i} \sim Ber(0.5)$, and are independent. We consider three different sample size scenarious $n = \{20, 50, 100\}$. Skewing shock matrix is sampled as $\mathbf{Z}_{2i} \sim MN(\mathbf{0}, \mathbf{I}, \mathbf{I})\mathbf{I}_{(0,\infty)}$ and matrix variate error $\mathbf{Z}_{1i} \sim MN(\mathbf{0}, \sigma^2\mathbf{R}_1, \mathbf{R}_2)$, where $\sigma^2 = 1$, and equicorrelation matrices $\mathbf{R}_1$, $\mathbf{R}_2$ with $\rho_1, \rho_2 = 0.6$ and 1 for diagonal elements.

Both methods (BTN/BSTN) converge rapidly regardless of initial values of parameters. BTN and BSTN both use the flat prior for $\pi(\mathcal{B}) \sim TN(\mathbf{0}, 10\mathbf{I}, 10\mathbf{I}, 10\mathbf{I})$, with a flat (non-informative) prior providing justification for valid posterior inference when considerable prior opinion is not available for $\mathcal{B}$. Performances of methods are evaluated from 30 data replications in terms of mean squared error (MSE), $\text{MSE}(\hat{\mathbf{B}}) = \sum_{q=1}^{Q}\sum_{j=1}^{J}(\hat{\mathbf{B}}_j^{(q)} - \mathbf{B}_j)^2/QJ$, where $\mathbf{B}_j$ is the jth element of the vectorized true parameter, with $\hat{\mathbf{B}}_j^{(q)}$ the posterior mean from 1,000 samples (keeping every 5th sample, after discarding the first 100 burn-in samples) from the qth simulated data, $q = 1, \ldots, Q$. We compare the OLS, ENV, and BTRR methods. While ENV used the estimated envelope dimensions, the rank of both $\mathcal{B} \in \mathbb{R}^{20\times6\times3}$ and $\mathcal{B} \in \mathbb{R}^{10\times3\times3}$ in the BTRR method was set to 3, given that the rank of margin for the PARAFAC decomposition of $\mathcal{B}$ cannot exceed 3.

For both scenarios, BTN and BSTN present better performances compared to OLS, ENV, and BTRR in all scenarios displayed in Table A.1 and A.2. BTN beats other estimators, especially when the simulated data assumed normality, while BSTN outperforms the alternatives for highly skewed data regardless of sample size. All estimated values of the coefficients gets closer to the true with increasing sample size. Furthermore, BSTN estimates the skewness level adequately in terms of MSE. On the contrary, the BTRR setup leads to highly sparse coefficients, and a poor performance. In summary, both BTN and BSTN exhibits superior performances compared to the alternatives, especially when the true coefficients are non-low rank with opposite signs.

Table A.1: Simulation study: The MSE (standard errors) of the estimated $\mathcal{B}$ and λ, obtained from fitting our BTN/BTSN and other competing methods (OLS, ENV and BTRR) to matrix-variate data generated under Model 1, across the 2 scenarios and sample sizes choices. The lowest MSE for each case is highlighted in boldface.

$\mathbb{R}^{10\times3\times3}$		BTN	BSTN		OLS	ENV	BTRR
Scenario	n	$\mathcal{B}$	$\mathcal{B}$	λ	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$
i	20	**0.1922** (0.071)	0.2016 (0.078)	0.0792 (0.037)	0.2100 (0.095)	0.2115 (0.075)	0.6291 (0.071)
	50	**0.0872** (0.031)	0.0892 (0.032)	0.0785 (0.041)	0.0901 (0.035)	0.0984 (0.025)	0.4972 (0.056)
	100	**0.0334** (0.012)	0.0340 (0.013)	0.0735 (0.035)	0.0338 (0.014)	0.0360 (0.011)	0.3662 (0.055)
ii	20	0.5398 (0.111)	**0.4976 (0.106)**	0.0108 (0.011)	0.5724 (0.154)	0.5832 (0.115)	0.7069 (0.051)
	50	0.2054 (0.051)	**0.1817 (0.045)**	0.0103 (0.012)	0.2095 (0.051)	0.2135 (0.053)	0.6299 (0.050)
	100	0.0907 (0.018)	**0.0827 (0.014)**	0.0065 (0.008)	0.0910 (0.019)	0.0934 (0.018)	0.5321 (0.046)
$\mathbb{R}^{20\times6\times3}$							
i	20	**0.2023 (0.053)**	0.2294 (0.078)	0.0131 (0.020)	0.2510 (0.095)	0.2637 (0.068)	0.6691 (0.034)
	50	**0.0689 (0.021)**	0.0717 (0.021)	0.0060 (0.009)	0.0751 (0.026)	0.0778 (0.024)	0.5260 (0.040)
	100	**0.0394 (0.009)**	0.0405 (0.010)	0.0049 (0.009)	0.0413 (0.010)	0.0433 (0.008)	0.4655 (0.026)
ii	20	0.5600 (0.104)	**0.5246 (0.094)**	0.0290 (0.014)	0.5971 (0.106)	0.6032 (0.095)	0.7833 (0.016)
	50	0.1996 (0.030)	**0.1720 (0.030)**	0.0244 (0.018)	0.2000 (0.030)	0.2044 (0.025)	0.6972 (0.038)
	100	0.0976 (0.010)	**0.0863 (0.010)**	0.0097 (0.009)	0.0990 (0.011)	0.1023 (0.009)	0.6972 (0.038)

Table A.2: Simulation study: The MSE (standard errors) of the estimated $\mathcal{B}$ and λ from using the BTN, BTSN and other competing methods (OLS, ENV and BTRR) to matrix-variate data simulated from two scenarios and under 3 different sample sizes (n). For each n and scenario combination, the lowest MSE among competing methods is highlighted in boldface.

$\mathbb{R}^{10\times3\times3}$		BTN	BSTN		OLS	ENV	BTRR
Scenario	n	$\mathcal{B}$	$\mathcal{B}$	λ	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$
i	20	**0.2032** (0.077)	0.2064 (0.089)	0.0773 (0.044)	0.2100 (0.095)	0.2285 (0.147)	1.2764 (0.210)
	50	**0.0899** (0.031)	0.0908 (0.034)	0.0747 (0.017)	0.0902 (0.035)	0.0910 (0.038)	1.0528 (0.157)
	100	**0.0331** (0.012)	0.0340 (0.013)	0.0692 (0.018)	0.0339 (0.014)	0.0348 (0.019)	0.9410 (0.087)
ii	20	0.5417 (0.109)	**0.4842** (0.154)	0.0123 (0.009)	0.5534 (0.118)	0.5931 (0.275)	1.4904 (0.128)
	50	0.2087 (0.052)	**0.1834** (0.049)	0.0086 (0.013)	0.2095 (0.051)	0.2604 (0.049)	1.2817 (0.108)
	100	0.0907 (0.018)	**0.0832** (0.016)	0.0062 (0.005)	0.0911 (0.020)	0.2144 (0.025)	1.1866 (0.084)
$\mathbb{R}^{20\times6\times3}$							
i	20	**0.2255** (0.068)	0.2480 (0.089)	0.0321 (0.026)	0.2510 (0.095)	0.2292 (0.093)	1.2742 (0.170)
	50	**0.0718** (0.024)	0.0759 (0.026)	0.0228 (0.008)	0.0751 (0.026)	0.0760 (0.027)	1.0712 (0.102)
	100	**0.0400** (0.009)	0.0413 (0.010)	0.0224 (0.004)	0.0414 (0.010)	0.0427 (0.014)	0.9996 (0.064)
ii	20	0.5721 (0.099)	**0.5436** (0.099)	0.0305 (0.019)	0.5997 (0.099)	0.6833 (0.118)	1.4689 (0.176)
	50	0.2016 (0.029)	**0.1743** (0.032)	0.0297 (0.018)	0.2000 (0.030)	0.3823 (0.037)	1.2709 (0.107)
	100	0.0978 (0.012)	**0.0862** (0.011)	0.0275 (0.013)	0.0990 (0.011)	0.2824 (0.027)	1.1868 (0.063)

To assess our model performance in light of existing estimators in image recovery problems, we consider additional simulation setup where we generate two-way tensor responses $\mathbf{Y}_i = \mathbf{B}x_i + \lambda|\mathbf{Z}_{2i}| + \mathbf{Z}_{1i}$ for $i = 1, \ldots, n(= 20)$, where $\mathbf{Y}_i \in \mathbb{R}^{32\times32}$, x_i is a scalar taking values 0 or 1 (indicative of disease and control groups, respectively), latent variable $\mathbf{Z}_{2i}$ follows matrix normal distribution,

$\mathbf{Z}_{2i} \sim MN(\mathbf{0}, \mathbf{I}_{32}, \mathbf{I}_{32})$, and the random noise $\mathbf{Z}_{1i}$ generated from a matrix normal distribution $MN(\mathbf{0}, \mathbf{I}_{32}, \mathbf{I}_{32})$. $\mathbf{B}$ is set as a 32×32 coefficient matrix, with the elements 0 (white region), or 1 (black region). The true coefficient matrices (true signals), represented by mouse and camel, are displayed in the first column of Figure A.1. We consider two scenarios, i.e., setting the true skewness $\lambda = 0$ (first and third row), or 1 (second and fourth rows). The true ranks of $\mathbf{B}$ from mouse and camel images are 14 and 21, respectively.

Figure A.1 displays the estimated coefficient matrices $\hat{\mathbf{B}}$, corresponding to the two scenarios and shapes. On the overall, BSTN outperforms other estimators under skewness, even when the sample size is fairly small. When the true signal is the mouse and $\lambda = 0$, BTN, BSTN, ENV, and OLS capture the signals adequately, however, the ENV, OLS, and BTRR present vague image recovery compared to BTN and BSTN under $\lambda = 1$. Note, ENV sets the estimated dimensions following their own algorithm, while BTRR places the true rank of each shape. The rank of true $\mathbf{B}$ for the camel is larger than the mouse, and as expected, the recovery of the camel images are substantially better from the BTN and BSTN methods, compared to the other methods when $\lambda = 1$.

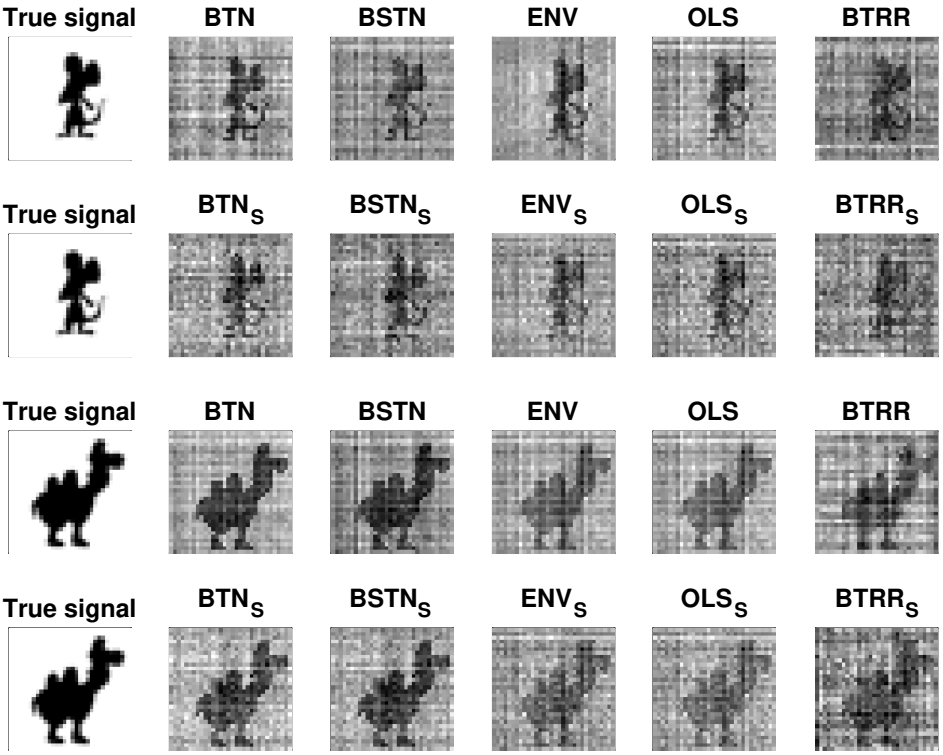

Figure A.1: Displayed are the true signals (Column 1), and recovered images (Columns 2-6) from fitting our BTN, BSTN, ENV, OLS and BTRR models to simulated 2-way tensor data (images) generated from Model 3, corresponding to the two scenarios and shapes. While the first and the third rows display estimated $\mathcal{B}$ under matrix normal data, the second and the fourth rows correspond to the skewed scenario, with the fitted models denoted by the suffix S.

A.6 GAAD Data Analysis: Additional Results

Table A.3: Average widths of the posterior interval estimates of covariate effects for all 28×6 teeth-site combinations, obtained from the BSTN and BTN fits with TN prior on $\mathcal{B}$, corresponding special cases with $\mathbf{R}_3 = \mathbf{I}$, and the OLS. The OLS fit uses 1000 bootstrap samples for variance estimation.

PPD	BTN	BSTN	BTN with $\mathbf{R}_3 = \mathbf{I}$	BSTN with $\mathbf{R}_3 = \mathbf{I}$	OLS
Intercept	1.4919	1.4241	1.5053	1.4580	1.4855
Age	0.0230	0.0216	0.0235	0.0219	0.0242
Gender	0.5875	0.5727	0.6141	0.5815	0.6693
BMI	0.5266	0.5163	0.5509	0.5245	0.5897
Smoker	0.5349	0.5232	0.5571	0.5316	0.5873
HbA1c	0.5513	0.5407	0.5574	0.5473	0.5700
CAL	BTN	BSTN	BTN with $\mathbf{R}_3 = \mathbf{I}$	BSTN with $\mathbf{R}_3 = \mathbf{I}$	OLS
Intercept	1.7536	1.7254	1.8315	1.7226	1.8292
Age	0.0276	0.0267	0.0289	0.0269	0.0294
Gender	0.7082	0.6860	0.7446	0.7015	0.8363
BMI	0.6470	0.6205	0.6701	0.6310	0.7481
Smoker	0.6479	0.6315	0.6794	0.6405	0.7401
HbA1c	0.6684	0.6481	0.7031	0.6630	0.7119

Table A.4: Fitting the BTN Model (with sparse tensor prior for $\mathcal{B}$) and the OLS to the GAAD data. Values in table are the posterior summaries of the overall covariate associations, and the skewness parameters, corresponding to the PPD (upperrow), and CAL (lower row)

BTN				OLS		
PPD	Median	SD	CI	PPD	Median	CI
Intercept	1.7860	0.1746	(1.3925, 1.8972)	Intercept	1.4885	(1.0745, 2.5630)
Age	0.0020	0.0024	(-0.0019, 0.0050)	Age	0.0015	(-0.0106, 0.0136)
Gender	-0.1195	0.0814	(-0.2332, 0.0408)	Gender	-0.1169	(-0.4573, 0.2120)
BMI	0.0427	0.0987	(-0.0647, 0.2540)	BMI	-0.0076	(-0.3081, 0.2816)
Smoker	0.1854	0.0811	(0.0553, 0.3287)	Smoker	0.1625	(-0.1261, 0.4612)
HbA1c	0.1702	0.1174	(0.1062, 0.4214)	HbA1c	0.2364	(-0.0504, 0.5196)
CAL	Median	SD	CI	CAL	Median	CI
Intercept	1.2487	0.1766	(0.8534, 1.3647)	Intercept	1.2422	(0.3413, 2.1705)
Age	0.0108	0.0022	(0.0066, 0.0128)	Age	0.0108	(-0.0040, 0.0254)
Gender	-0.2316	0.0976	(-0.3937, -0.0682)	Gender	-0.2219	(-0.6489, 0.1874)
BMI	0.0606	0.0737	(-0.0524, 0.1851)	BMI	0.0083	(-0.3737, 0.3744)
Smoker	0.3791	0.0738	(0.2161, 0.5040)	Smoker	0.3210	(-0.0379, 0.7022)
HbA1c	0.3781	0.1336	(0.1848, 0.5470)	HbA1c	0.3177	(-0.0401, 0.6718)

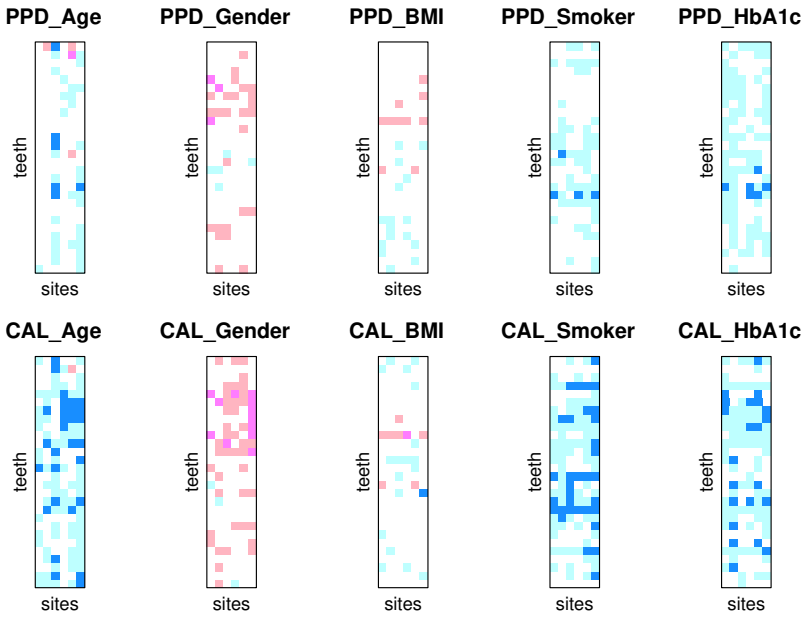

Figure A.2: Fitting the BTN Model with tensor normal prior on $\mathcal{B}$ to the GAAD data. Plotted are the D statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

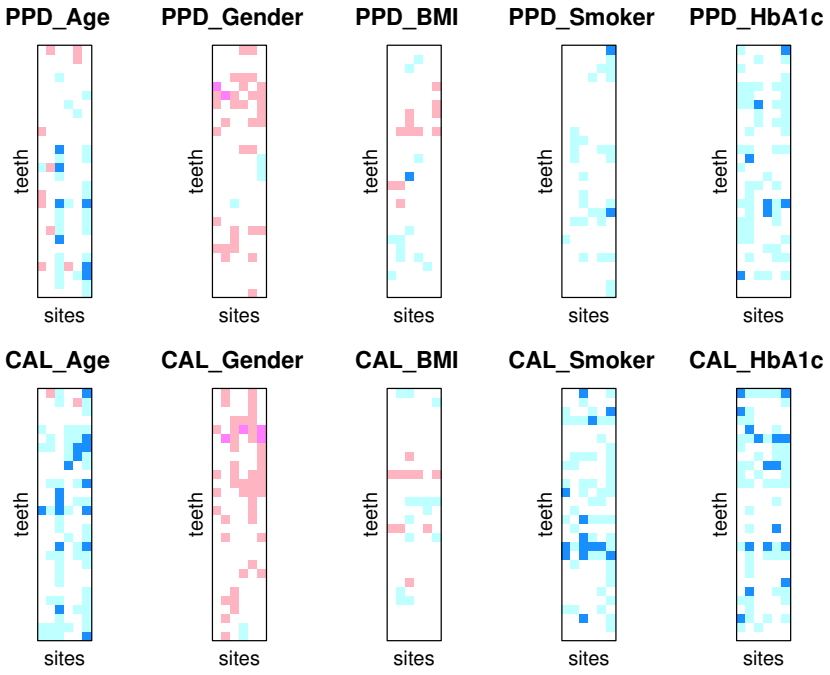

Figure A.3: Fitting the OLS model to the GAAD data, using 1000 bootstrap samples for variance estimation. Plotted are the D statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

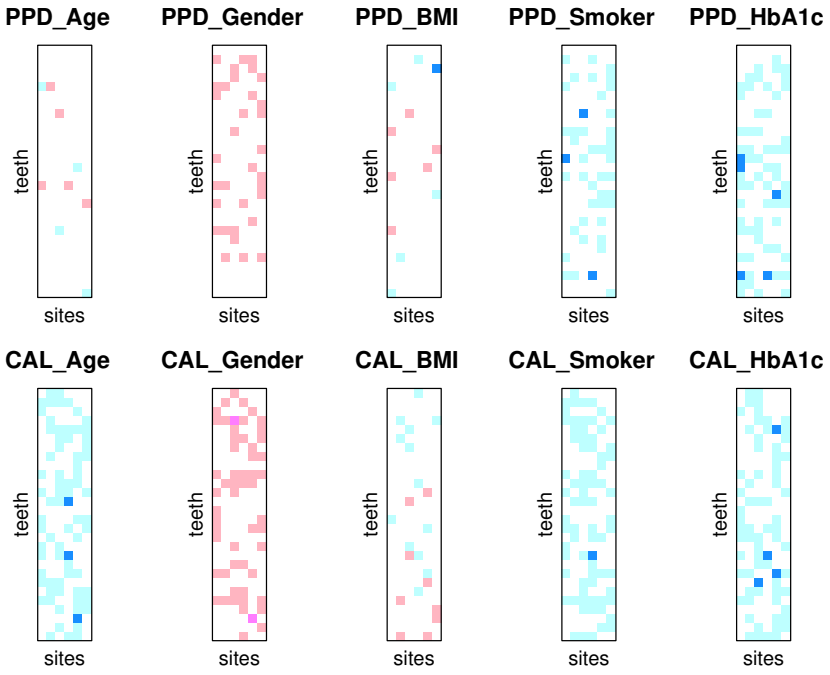

Figure A.4: Fitting the ENV model to the GAAD data, using 1000 bootstrap samples for variance estimation. Plotted are the D statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

APPENDIX B

APPENDIX OF CHAPTER 3

B.1 Technical Results and Proofs

Lemma 8. *Let* $\mathcal{S} \sim TN(0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, $\mathcal{Z} \sim TN(0, \boldsymbol{I}_{d_1}, \ldots, \boldsymbol{I}_{d_K})$, *and* $\mathcal{C}$ *denotes constant tensor, then* $\mathrm{E}_{\mathcal{S}}\{P(\mathcal{Z} \leq \mathcal{C} + [\![\mathcal{S}; \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K]\!])\}$ *is equivalent to* $P(\mathcal{Z} \leq [\![\mathcal{C}; (\boldsymbol{I}_{d_1} + \boldsymbol{Q}_1\boldsymbol{\Sigma}_1\boldsymbol{Q}_1^\intercal)^{-\frac{1}{2}}, \ldots, (\boldsymbol{I}_{d_K} + \boldsymbol{Q}_K\boldsymbol{\Sigma}_K\boldsymbol{Q}_K^\intercal)^{-\frac{1}{2}}]\!])$.

Proof of Lemma 1.2 We firstly derive the multivariate case of Lemma 8. Let $\mathbf{s}_1 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_1)$, $\mathbf{z}_1 \sim N(\mathbf{0}, \mathbf{I}_{d_1})$, $\mathbf{c}_1 \in \mathbb{R}^{d_1}$ is a vector of constants, $\mathbf{Q}_1 \in \mathbb{R}^{d_1 \times d_1}$, and define $\boldsymbol{\Phi}_{d_1}$ is a cumulative density function of $\mathbf{z}_1$. Then, $E(\boldsymbol{\Phi}_{d_1}(\mathbf{c}_1 + \mathbf{Q}_1\mathbf{s}_1)) = E[P(\mathbf{z}_1 \leq \mathbf{c}_1 + \mathbf{Q}_1\mathbf{s}_1 | \mathbf{s}_1)] = P(\mathbf{z}_1 - \mathbf{Q}_1\mathbf{s}_1 \leq \mathbf{c}_1)$. Since the linear combination of multivariate normal distributions are still multivariate normal distribution, $\mathbf{z}_1 - \mathbf{Q}_1\mathbf{s}_1 \sim N(\mathbf{0}, \mathbf{I}_{d_1} + \mathbf{Q}_1^\intercal\boldsymbol{\Sigma}_1\mathbf{Q}_1)$. Hence, $E(\boldsymbol{\Phi}_{d_1}(\mathbf{c}_1 + \mathbf{Q}_1\mathbf{s}_1)) = \boldsymbol{\Phi}_{d_1}(\{\mathbf{I}_{d_1} + \mathbf{Q}_1^\intercal\boldsymbol{\Sigma}_1\mathbf{Q}_1\}^{-\frac{1}{2}}\mathbf{c}_1) = P(\mathbf{z}_1 \leq \{\mathbf{I}_{d_1} + \mathbf{Q}_1^\intercal\boldsymbol{\Sigma}_1\mathbf{Q}_1\}^{-\frac{1}{2}}\mathbf{c}_1)$. We directly extend this proof to tensor case using Lemma 7. We extend d_1-dimensional vectors, $\mathbf{s}_1, \mathbf{z}_1, \mathbf{c}_1$ to $(d_1 \times \cdots \times d_K)$-dimensional tensors, $\mathcal{S}, \mathcal{Z}, \mathcal{C}$, respectively. Using the proof of multivariate case and Lemma 7, $E[P(\mathcal{Z} \leq \mathcal{C} + [\![\mathcal{S}; \mathbf{Q}_1, \ldots, \mathbf{Q}_K]\!])] = \boldsymbol{\Psi}_{d_1,\ldots,d_K}(\mathcal{C} \times_1 \{\mathbf{I}_{d_1} + \mathbf{Q}_1^\intercal\boldsymbol{\Sigma}_1\mathbf{Q}_1\} \cdots \times_K \{\mathbf{I}_{d_K} + \mathbf{Q}_K^\intercal\boldsymbol{\Sigma}_K\mathbf{Q}_K\}) = P(\mathcal{Z} \leq [\![\mathcal{C}; (\mathbf{I}_{d_1} + \mathbf{Q}_1\boldsymbol{\Sigma}_1\mathbf{Q}_1^\intercal)^{-\frac{1}{2}}, \ldots, (\mathbf{I}_{d_K} + \mathbf{Q}_K\boldsymbol{\Sigma}_K\mathbf{Q}_K^\intercal)^{-\frac{1}{2}}]\!])$. This completes the proof. □

Lemma 9. *The mgf of tensor skewed normal distribution,* $M_{\mathcal{Y}}(\mathcal{T})$ *is given by*

$$M_{\mathcal{Y}}(\mathcal{T}) = 2^{(d_1 \times \cdots \times d_K)} P(\mathcal{Z} \leq [\![\mathcal{T}; \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K]\!]) \times \exp\left\{\frac{1}{2}\left\langle [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!], [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!] \right\rangle + \langle \mathcal{M}, \mathcal{T} \rangle\right\},$$

based Lemma 8.

Proof of Lemma 9

$$M_{\mathcal{Y}}(\mathcal{T}) \quad = M_{\mathcal{A}}(\mathcal{T}) \exp(\langle \mathcal{M}, \mathcal{T} \rangle), \text{ where } \mathcal{A} \sim STN(0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K; \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K).$$

Let $\boldsymbol{G}_1 = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Lambda}_1^2)^{-1}, \ldots, \boldsymbol{G}_K = (\boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^2)^{-1}$, $\boldsymbol{H}_1 = \boldsymbol{I}_{d_1} - \boldsymbol{\Lambda}_1 \boldsymbol{G}_1 \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{H}_K = (\boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^2)^{-1}$, *and* $\mathcal{Z} \sim TN(0; \boldsymbol{I}_{d_1}, \ldots, \boldsymbol{I}_{d_K})$.

$$
\begin{aligned}
&M_{\mathcal{A}}(\mathcal{T}) = 2^{(d_1 \times \cdots \times d_K)} \int_{\mathbb{R}^{d_1}} \cdots \int_{\mathbb{R}^{d_K}} \prod_{k=1}^{K} \det(\boldsymbol{G}_k)^{(d_1 \times \cdots \times d_K)/(2d_k)} (2\pi)^{-(d_1 \times \cdots \times d_K)/2} \\
&\times \exp\left\{ -\frac{1}{2} \left\langle [\![\mathcal{A}; \boldsymbol{G}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{-\frac{1}{2}}]\!], [\![\mathcal{A}; \boldsymbol{G}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{-\frac{1}{2}}]\!] \right\rangle + \langle \mathcal{A}, \mathcal{T} \rangle \right\} \\
&\times P(\mathcal{A} \leq [\![\mathcal{A}; \boldsymbol{H}_1^{-\frac{1}{2}} \boldsymbol{\Lambda}_1 \boldsymbol{G}_1, \ldots, \boldsymbol{H}_K^{-\frac{1}{2}} \boldsymbol{\Lambda}_K \boldsymbol{G}_K]\!]) d\mathcal{A}_{i_1 \cdots i_{K-1}:} \cdots d\mathcal{A}_{:i_2 \cdots i_K} \\
&= 2^{(d_1 \times \cdots \times d_K)} \exp\left\{ \frac{1}{2} \left\langle [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!], [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!] \right\rangle \right\} \\
&\times \int_{\mathbb{R}^{d_1}} \cdots \int_{\mathbb{R}^{d_K}} \prod_{k=1}^{K} \det(\boldsymbol{G}_k)^{(d_1 \times \cdots \times d_K)/(2d_k)} (2\pi)^{-(d_1 \times \cdots \times d_K)/2} \\
&\times \exp\left\{ -\frac{1}{2} \left\langle [\![\mathcal{A} - [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!]; \boldsymbol{G}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{-\frac{1}{2}}]\!], [\![\mathcal{A} - [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!]; \boldsymbol{G}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{-\frac{1}{2}}]\!] \right\rangle \right\} \\
&\times P(\mathcal{A} \leq [\![\mathcal{A}; \boldsymbol{H}_1^{-\frac{1}{2}} \boldsymbol{\Lambda}_1 \boldsymbol{G}_1, \ldots, \boldsymbol{H}_K^{-\frac{1}{2}} \boldsymbol{\Lambda}_K \boldsymbol{G}_K]\!]) d\mathcal{A}_{i_1 \cdots i_{K-1}:} \cdots d\mathcal{A}_{:i_2 \cdots i_K} \\
&\text{Let } \mathcal{Z} = \mathcal{A} - [\![\mathcal{T}; \mathbf{G}_1^{\frac{1}{2}}, \ldots, \mathbf{G}_K^{\frac{1}{2}}]\!], \text{ then} \\
&M_{\mathcal{A}}(\mathcal{T}) = 2^{(d_1 \times \cdots \times d_K)} \exp\left\{ \frac{1}{2} \left\langle [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!], [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!] \right\rangle \right\} \\
&\times \int_{\mathbb{R}^{d_1}} \cdots \int_{\mathbb{R}^{d_K}} \prod_{k=1}^{K} \det(\boldsymbol{G}_k)^{(d_1 \times \cdots \times d_K)/(2d_k)} (2\pi)^{-(d_1 \times \cdots \times d_K)/2} \\
&\times \exp\left\{ -\frac{1}{2} \left\langle [\![\mathcal{Z}; \boldsymbol{G}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{-\frac{1}{2}}]\!], [\![\mathcal{Z}; \boldsymbol{G}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{-\frac{1}{2}}]\!] \right\rangle \right\} \\
&\times P(\mathcal{Z} \leq [\![\mathcal{Z} + [\![\mathcal{T}; \boldsymbol{G}_1, \ldots, \boldsymbol{G}_K]\!]; \boldsymbol{H}_1^{-\frac{1}{2}} \boldsymbol{\Lambda}_1 \boldsymbol{G}_1, \ldots, \boldsymbol{H}_K^{-\frac{1}{2}} \boldsymbol{\Lambda}_K \boldsymbol{G}_K]\!]) d\mathcal{Z}_{i_1 \cdots i_{K-1}:} \cdots d\mathcal{Z}_{:i_2 \cdots i_K} \\
&= 2^{(d_1 \times \cdots \times d_K)} \exp\left\{ \frac{1}{2} \left\langle [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!], [\![\mathcal{T}; \boldsymbol{G}_1^{\frac{1}{2}}, \ldots, \boldsymbol{G}_K^{\frac{1}{2}}]\!] \right\rangle \right\} P(\mathcal{Z} \leq [\![\mathcal{T}; \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K]\!]) \quad \square
\end{aligned}
$$

.

Lemma 10. *Partition* $\mathcal{E} = \begin{pmatrix} \mathcal{E}^{(1)} \\ \mathcal{E}^{(2)} \end{pmatrix} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $0 = \begin{pmatrix} 0^{(1)} \\ 0^{(2)} \end{pmatrix}$, *and* $\mathcal{Z}_2 = \begin{pmatrix} \mathcal{Z}_2^{(1)} \\ \mathcal{Z}_2^{(2)} \end{pmatrix}$ *along* K*th mode, let* $d_K = 2$, $\boldsymbol{D}_{\boldsymbol{\sigma}}^2 = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, $\boldsymbol{\Lambda}_K = \mathrm{diag}(\lambda_1, \lambda_2)$, $\boldsymbol{R}_K = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, *and* $\boldsymbol{\Sigma}_K = \boldsymbol{D}_{\boldsymbol{\sigma}} \boldsymbol{R}_K \boldsymbol{D}_{\boldsymbol{\sigma}} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. *Here, particularly, we derive the first subset of components of* $\mathcal{E}$ *along* K*th mode. The marginal density of* $\mathcal{E}^{(1)}$ *would be* $f\left(\mathcal{E}^{(1)}\right) = 2^{(d_1 \ldots d_{K-1})} \prod_{k=1}^{K-1} \det(\boldsymbol{Q}_k)^{-\frac{(d_1 \times \cdots \times d_{K-1})}{(2d_k)}} \det(\sigma_1^2 + \lambda_1^2)^{-\frac{1}{2}} \times$

$$\left[1+\left\langle [\![\mathcal{E}^{(1)}; \boldsymbol{Q}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{Q}_{K-1}^{-\frac{1}{2}}, (\sigma_1^2+\lambda_1^2)^{-\frac{1}{2}}]\!], [\![\mathcal{E}^{(1)}; \boldsymbol{Q}_1^{-\frac{1}{2}}, \ldots, \boldsymbol{Q}_{K-1}^{-\frac{1}{2}}, (\sigma_1^2+\lambda_1^2)^{-\frac{1}{2}}]\!] \right\rangle \Big/ \nu \right]^{-\frac{\nu+2(d_1\times\cdots\times d_K)}{2}}$$

$\times P(\mathcal{Z}_2^{(1)} > 0)$, *where* $\mathcal{Z}_2^{(1)} \sim TEl\left(\lambda_1(\sigma_1^2+\lambda_1^2)^{-1}\mathcal{E}^{(1)}; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{K-1}, \sigma_1^2-\lambda_1(\sigma_1^2+\lambda_1^2)^{-1}\lambda_1; g^{(d_1\ldots d_{K-1})}\right)$.

Proof of Lemma 10 Take $g^{d_1,\ldots,d_{K-1}}(u) = (1+\frac{u}{\nu})^{-\frac{\nu+(d_1\times\cdots\times d_K)}{2}}$ in Proposition 1 in the main text. □

Lemma 11. *Cumulative conditional density of* $\mathrm{vec}(\mathcal{Z}_2)$ *is*

$$t\left(\mathrm{vec}(\mathcal{Z}_2)\middle|(\boldsymbol{\Lambda}_K\boldsymbol{Q}_K^{-1}\otimes\cdots\otimes\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^{-1})\boldsymbol{y}^*, \frac{\nu+\boldsymbol{y}^{*\intercal}(\boldsymbol{Q}_K^{-1}\otimes\cdots\otimes\boldsymbol{Q}_1^{-1})\boldsymbol{y}^*}{\nu+\prod_{k=1}^K d_k}, \nu+\textstyle\prod_{k=1}^K d_k\right).$$

Proof of Lemma 11

From vectorized (3.4) in the main text and Lemma A.1 in [48],

$$g_a^{(\prod_{k=1}^K d_k)}(u;\nu) = \frac{\Gamma(\frac{\prod_{k=1}^K d_k}{2})}{\pi^{\frac{\prod_{k=1}^K d_k}{2}}} \frac{g(a+u; 2\prod_{k=1}^K d_k, \nu)}{\int_0^\infty r^{\prod_{k=1}^K d_k/2-1} g(a+u; 2\prod_{k=1}^K d_k, \nu)dr}$$

$$= \frac{\Gamma(\frac{\prod_{k=1}^K d_k}{2})}{\{\pi+(\nu+\prod_{k=1}^K d_k)\}^{\frac{\prod_{k=1}^K d_k}{2}}} \left(\frac{\nu+\prod_{k=1}^K d_k}{\nu+a}\right)^{\prod_{k=1}^K d_k} \left(1+\frac{u}{\nu+\prod_{k=1}^K d_k}\frac{\nu+\prod_{k=1}^K d_k}{\nu+1}\right)^{-\frac{\nu+2\prod_{k=1}^K d_k}{2}}$$

Then, the cumulative conditional density is

$$t\left(\mathrm{vec}(\mathcal{Z}_2)\middle|(\boldsymbol{\Lambda}_K\mathbf{Q}_K^{-1}\otimes\cdots\otimes\boldsymbol{\Lambda}_1\mathbf{Q}_1^{-1})\mathbf{y}^*, \frac{\nu+\mathbf{y}^{*\intercal}(\mathbf{Q}_K^{-1}\otimes\cdots\otimes\mathbf{Q}_1^{-1})\mathbf{y}^*}{\nu+\prod_{k=1}^K d_k}, \nu+\prod_{k=1}^K d_k\right)$$

where $\mathbf{Q}_k = \boldsymbol{\Sigma}_k + \boldsymbol{\Lambda}_k^\intercal\boldsymbol{\Lambda}_k$ for $k=1,\ldots,K-1$, $\mathbf{Q}_K = \boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^\intercal\mathbf{D}_{\boldsymbol{\sigma}}^2\boldsymbol{\Lambda}_K$, $\mathbf{y}^* = \mathrm{vec}(\mathcal{Y}) - \mathrm{vec}(\mathcal{M})$. □

Proof of Theorem 2 Theorem 2 is extended from Theorem 2.16 of [18] in the context of tensor responses. Since $E[\mathcal{M}+\mathcal{Z}_2\times_1\boldsymbol{\Lambda}_1\times_2\boldsymbol{\Lambda}_2\cdots\times_K\boldsymbol{\Lambda}_K] = \mathcal{M}+E[\mathcal{Z}_2]\times_1\boldsymbol{\Lambda}_1\times_2\boldsymbol{\Lambda}_2\cdots\times_K\boldsymbol{\Lambda}_K = \mathcal{M}$ which follows $E[\mathcal{Z}_2] = \mathbf{0}$ due to the assumption that $\mathcal{Z}_2 \sim TEL(\mathbf{0}, \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}_{\boldsymbol{\sigma}}^2; g^{(d_1,\ldots,d_K)})$. Let $\mathcal{S} = \mathcal{M}+\mathcal{Z}_2\times_1\boldsymbol{\Lambda}_1\times_2\boldsymbol{\Lambda}_2\cdots\times_K\boldsymbol{\Lambda}_K$. We need to show that

$$\begin{aligned}
&\mathbf{S}_{(1)} \sim MEL\left(\mathbf{M}_{(1)}; \mathbf{V}_{d_1}, \mathbf{V}_{d_2}\otimes\cdots\otimes\mathbf{V}_{d_K}; g^{(d_1, \prod_{k=2}^K d_k)}\right)\\
&\vdots\\
&\mathbf{S}_{(K-1)} \sim MEL\left(\mathbf{M}_{(K-1)}; \mathbf{V}_{d_{K-1}}, \mathbf{V}_{d_1}\otimes\cdots\otimes\mathbf{V}_{d_K}; g^{(d_{K-1}, d_K\prod_{k=1}^{K-2} d_k)}\right)\\
&\mathbf{S}_{(K)} \sim MEL\left(\mathbf{M}_{(K)}; \mathbf{V}_K, \mathbf{V}_{d_1}\otimes\cdots\otimes\mathbf{V}_{d_{K-1}}; g^{(d_K, \prod_{k=1}^{K-1} d_k)}\right),
\end{aligned}$$

where 'MEL' denotes matrix-variate elliptical distribution. Among above statement, we only need to show the first equation holds and the other $(K-1)$ cases are dealt with the same way. By the definition of $\mathcal{S}$, We have

$$\mathbf{S}_{(1)} = \mathbf{M}_{(1)} + \mathbf{\Lambda}_1^\intercal \mathbf{Z}_{2(1)}(\mathbf{\Lambda}_2 \otimes \cdots \otimes \mathbf{\Lambda}_K).$$

Let $\mathbf{C} \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{V}_1 = \mathbf{\Lambda}_1 \mathbf{I}_{d_1} \mathbf{\Lambda}_1^\intercal = \mathbf{\Lambda}_1 \mathbf{\Lambda}_1^\intercal$ and $\mathbf{V}_2 = \mathbf{\Lambda}_2 \mathbf{I}_{d_2} \mathbf{\Lambda}_2^\intercal = \mathbf{\Lambda}_2 \mathbf{\Lambda}_2^\intercal$. Then $\mathbf{C} \sim MEL(\mathbf{0}; \mathbf{V}_1, \mathbf{V}_2; g^{(d_1,d_2)})$ if and only if there exists a $\mathbf{G} \sim MEL(\mathbf{0}, \mathbf{I}_{d_1}, \mathbf{I}_{d_2}; g^{(d_1,d_2)})$ such that $\mathbf{C} = \mathbf{\Lambda}_1 \mathbf{G} \mathbf{\Lambda}_2^\intercal$. Then, $\mathbf{S}_{(1)} \sim MEL(\mathbf{M}_{(1)}; \mathbf{V}_1, \mathbf{V}_2 \otimes \cdots \otimes \mathbf{V}_K; g^{(d_1, \prod_{k=2}^K d_k)}))$ since $\mathbf{V}_1 = \mathbf{\Lambda}_1 \mathbf{\Lambda}_1^\intercal$ and all the rows of $\mathbf{S}_{(1)}$ denoted by $\mathbf{S}_{(1):}$ follows $EL(\mathbf{M}_{(1):}; \mathbf{V}_2 \otimes \cdots \otimes \mathbf{V}_K; g^{\prod_{k=2}^K d_k})$ by the property of Kronecker product, and 'EL' denotes multivariate elliptical distribution. Similarly we can show that

$$\mathbf{S}_{(K-1)} \sim MEL(\mathbf{M}_{(K-1)}; \mathbf{V}_{K-1}, \mathbf{V}_1 \otimes \cdots \otimes \mathbf{V}_{K-2} \otimes \mathbf{V}_K; g^{(d_{K-1} d_K \prod_{k=1}^{K-2} d_k)})$$
$$\mathbf{S}_{(K)} \sim MEL(\mathbf{M}_{(K)}; \mathbf{V}_K, \mathbf{V}_1 \otimes \cdots \otimes \mathbf{V}_{K-1}; g^{(d_K, \prod_{k=1}^{K-1} d_k)}),$$

which concludes the proof. □

Proof of Theorem 3 Let $\mathbf{A}_{k,2}$ be a matrix such that $\mathbf{A}_{k,2}^\intercal \mathbf{A}_{k,2} = \mathbf{\Sigma}_{k,22}$, $\mathbf{A}_{k,2} > \mathbf{0}$. For $k = 1, \ldots, K$,

$$\mathbf{\Sigma}_k = \mathbf{A}_k^\intercal \mathbf{A}_k = \begin{pmatrix} \mathbf{I} & \mathbf{\Sigma}_{k,12}\mathbf{\Sigma}_{k,22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_{k,11.2} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{k,22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{k,2}\mathbf{\Sigma}_{k,22}^{-1}\mathbf{A}_{k,21} & \mathbf{A}_{k,2} \end{pmatrix},$$

where $\mathbf{A}_k = \begin{pmatrix} \mathbf{A}_{k,11.2} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{k,2} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Sigma}_{k,22}^{-1}\mathbf{\Sigma}_{k,21} & \mathbf{I} \end{pmatrix}$, and $\mathbf{\Sigma}_{k,11.2} = \mathbf{A}_{k,11.2}^\intercal \mathbf{A}_{k,11.2}$.

Let $\mathcal{Z} \sim TEL(0, \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_K}; g^{(d_1,\ldots,d_K)})$, then, we have $\mathcal{Y} \overset{d}{=} \mathcal{M} + [\![\mathcal{Z}; \mathbf{A}_1^\intercal, \ldots, \mathbf{A}_K^\intercal]\!]$.

So, applying Theorem 2, $\mathcal{Y} \sim TEL(M, \mathbf{A}_1^\intercal \mathbf{A}_1, \ldots, \mathbf{A}_K^\intercal \mathbf{A}_K; g^{(d_1,\ldots,d_K)})$.

By abuse of notation, partition $\mathcal{M} = \begin{pmatrix} \mathcal{M}^{(1)} \\ \mathcal{M}^{(2)} \end{pmatrix}, \mathcal{Z} = \begin{pmatrix} \mathcal{Z}^{(1)} \\ \mathcal{Z}^{(2)} \end{pmatrix}$,

and applying Theorem 2 in the main text and Lemma 7, then,

$$\mathcal{Y} \overset{d}{=} \mathcal{M} + [\![\mathcal{Z}; \mathbf{A}_1^\intercal, \ldots, \mathbf{A}_K^\intercal]\!]$$
$$= \begin{pmatrix} \mathcal{M}^{(1)} \\ \mathcal{M}^{(2)} \end{pmatrix} + \begin{pmatrix} [\![\mathcal{Z}^{(1)}; \mathbf{A}_{1,11.2}^\intercal, \ldots, \mathbf{A}_{K,11.2}^\intercal]\!] + [\![\mathcal{Z}^{(2)}; \boldsymbol{\Sigma}_{1,12}\boldsymbol{\Sigma}_{1,22}^{-1}\mathbf{A}_{1,2}^\intercal, \ldots, \boldsymbol{\Sigma}_{K,12}\boldsymbol{\Sigma}_{K,22}^{-1}\mathbf{A}_{K,2}^\intercal]\!] \\ [\![\mathcal{Z}^{(2)}; \mathbf{A}_{1,2}^\intercal, \ldots, \mathbf{A}_{K,2}^\intercal]\!] \end{pmatrix}$$

and

$$\mathcal{Y}^{(1)}|\mathcal{Y}^{(2)} = [\![\mathcal{Z}^{(1)}; \mathbf{A}_{1,11.2}^\intercal, \ldots, \mathbf{A}_{K,11.2}^\intercal]\!] + [\![\mathcal{Z}^{(2)}; \boldsymbol{\Sigma}_{1,12}\boldsymbol{\Sigma}_{1,22}^{-1}\mathbf{A}_{1,2}^\intercal, \ldots, \boldsymbol{\Sigma}_{K,12}\boldsymbol{\Sigma}_{K,22}^{-1}\mathbf{A}_{K,2}^\intercal]\!] + \mathcal{M}^{(1)}$$
$$|\mathcal{M}^{(2)} + [\![\mathcal{Z}^{(2)}; \mathbf{A}_{1,2}^\intercal, \ldots, \mathbf{A}_{K,2}^\intercal]\!]$$
$$= \mathcal{M}^{(1)} + [\![\mathcal{Y}^{(2)} - \mathcal{M}^{(2)}; \boldsymbol{\Sigma}_{1,12}\boldsymbol{\Sigma}_{1,22}^{-1}, \ldots, \boldsymbol{\Sigma}_{K,12}\boldsymbol{\Sigma}_{K,22}^{-1}]\!] + [\![\mathcal{Z}^{(1)}; \mathbf{A}_{1,11.2}^\intercal, \ldots, \mathbf{A}_{K,11.2}^\intercal]\!]$$
$$|\mathcal{Z}^{(2)} = [\![\mathcal{Y}^{(2)} - \mathcal{M}^{(2)}; \mathbf{A}_{1,2}^{-1}, \ldots, \mathbf{A}_{K,2}^{-1}]\!], \text{ where } \mathbf{A}_{k,2}^{-1} = \boldsymbol{\Sigma}_{k,2}^{-\frac{1}{2}} \text{ for } k = 1, \ldots, K.$$

The density generating function for $\mathcal{Y}^{(1)}|\mathcal{Y}^{(2)}$ is simple extension of Lemma A.1 in [48], so it should be

$$g_a^{(d_{1,1},\ldots,d_{K,1})}(u) = \frac{\Gamma(d_{1,1} \times \cdots \times d_{K,1}/2)}{\pi^{(d_{1,1}\times\cdots\times d_{K,1})/2}} \frac{g(a+u; d_1, \ldots, d_K)}{\int_0^\infty r^{(d_{1,1}\times\cdots\times d_{K,1})/2-1} dr}$$

for some $r \geq 0$ which is independent of u, here $u = \mathcal{Y}^{(1)}$, and $a = q(\mathcal{Y}^{(2)}) = \langle \mathcal{Z}^{(2)}, \mathcal{Z}^{(2)} \rangle$
$$= \left\langle \left[\!\!\left[\mathcal{Y}^{(2)} - \mathcal{M}^{(2)}; \boldsymbol{\Sigma}_{1,22}^{-\frac{1}{2}}, \ldots, \boldsymbol{\Sigma}_{K,22}^{-\frac{1}{2}} \right]\!\!\right], \left[\!\!\left[\mathcal{Y}^{(2)} - \mathcal{M}^{(2)}; \boldsymbol{\Sigma}_{1,22}^{-\frac{1}{2}}, \ldots, \boldsymbol{\Sigma}_{K,22}^{-\frac{1}{2}} \right]\!\!\right] \right\rangle. \qquad \square$$

Proof of Theorem 4

Consider the transformation of tensor variables along each mode.

$$\begin{pmatrix} \mathcal{E} \\ \mathcal{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \end{pmatrix} \times_1 \begin{pmatrix} \mathbf{I}_{d_1} & \boldsymbol{\Lambda}_1 \\ \mathbf{0} & \mathbf{I}_{d_1} \end{pmatrix} \times_2 \begin{pmatrix} \mathbf{I}_{d_2} & \boldsymbol{\Lambda}_2 \\ \mathbf{0} & \mathbf{I}_{d_2} \end{pmatrix} \cdots \times_K \begin{pmatrix} \mathbf{I}_{d_K} & \boldsymbol{\Lambda}_K \\ \mathbf{0} & \mathbf{I}_{d_K} \end{pmatrix}$$
$$= \left[\!\!\left[\begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \end{pmatrix}; \mathbf{H}_1, \ldots, \mathbf{H}_K \right]\!\!\right], \text{ where } \mathbf{H}_k = \begin{pmatrix} \mathbf{I}_{d_k} & \boldsymbol{\Lambda}_k \\ \mathbf{0} & \mathbf{I}_{d_k} \end{pmatrix} \text{ for } k = 1, \ldots, K, \text{ and } \mathbf{0} \text{ is the null matrix.}$$

By applying Proposition 1, we observe that

$$\begin{pmatrix} \mathcal{E} \\ \mathcal{Z}_2 \end{pmatrix} \sim TEL\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 + \boldsymbol{\Lambda}_1^\intercal \boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_1 \\ \boldsymbol{\Lambda}_1 & \mathbf{I}_{d_1} \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^\intercal \mathbf{D}_{\boldsymbol{\sigma}}^2 \boldsymbol{\Lambda}_K & \boldsymbol{\Lambda}_K \\ \boldsymbol{\Lambda}_K & \mathbf{D}_{\boldsymbol{\sigma}}^2 \end{pmatrix}; g^{(2d_1,\ldots,2d_K)} \right).$$

Our goal is to obtain the conditional density of $f(\mathcal{E}|\mathcal{Z}_2 > 0)$ from the above joint distribution. Applying Bayes Theorem, we have

$$f(\mathcal{E}|\mathcal{Z}_2 > 0) = \frac{f(\mathcal{E}) P(\mathcal{Z}_2 > 0|\mathcal{E})}{P(\mathcal{Z}_2 > 0)}$$

Since $\mathcal{Z}_2 \sim TEL(0, \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}^2_{\boldsymbol{\sigma}}; g^{(d_1,\ldots,d_K)})$ marginally, $P(\mathcal{Z}_2 > 0) = 2^{-(d_1 \times \cdots \times d_K)}$. Then, the conditional density $\mathcal{E}|\mathcal{Z}_2 > 0$ is given by

$$f(\mathcal{E}|\mathcal{Z}_2 > 0) = 2^{(d_1 \times \cdots \times d_K)} f_{\mathcal{E}}(\mathcal{E}|\mathbf{0}, \boldsymbol{\Sigma}_1 + \boldsymbol{\Lambda}_1^\intercal \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^\intercal \mathbf{D}^2_{\boldsymbol{\sigma}} \boldsymbol{\Lambda}_K; g^{(d_1,\ldots,d_K)}) P(\mathcal{Z}_2 > 0|\mathcal{E}).$$

Using Theorem 3, we have the conditional density $f(\mathcal{Z}_2 > 0|\mathcal{E})$ as

$$\mathcal{Z}_2|\mathcal{E} \sim TEl\Big(\big[\!\big[\mathcal{E}; \boldsymbol{\Lambda}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Lambda}_1^\intercal \boldsymbol{\Lambda}_1)^{-1}, \ldots, \boldsymbol{\Lambda}_K(\boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^\intercal \mathbf{D}^2_{\boldsymbol{\sigma}} \boldsymbol{\Lambda}_K)^{-1}\big]\!\big]; \\ \mathbf{I}_{d_1} - \boldsymbol{\Lambda}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Lambda}_1^\intercal \boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\Lambda}_1, \ldots, \mathbf{D}^2_{\boldsymbol{\sigma}} - \boldsymbol{\Lambda}_K(\boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^\intercal \mathbf{D}^2_{\boldsymbol{\sigma}} \boldsymbol{\Lambda}_K)^{-1}\boldsymbol{\Lambda}_K; g^{(d_1,\ldots,d_K)}_{q(\mathcal{E})} \Big),$$

where $q(\mathcal{E}) = \big[\!\big[\mathcal{E}; \boldsymbol{\Sigma}_1 + \boldsymbol{\Lambda}_1^\intercal \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Sigma}_K + \boldsymbol{\Lambda}_K^\intercal \mathbf{D}^2_{\boldsymbol{\sigma}} \boldsymbol{\Lambda}_K\big]\!\big]$. Hence the proof is complete. □

Proof of Proposition 1

Apply the arguments in the proofs of Theorems 2 and 4. □

Proof of Theorem 5

Note that the $(d_1 \times \cdots \times d_K)$-dimensional tensor normal (TN) distributions,

$$\mathcal{Y}|\mathcal{Z}_2 \sim TN(\mathcal{M} + \big[\!\big[\mathcal{Z}_2; \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K\big]\!\big], \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K/\delta)$$

$$\mathcal{Z}_2 \sim TN(0, \mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_{K-1}}, \mathbf{D}^2_{\boldsymbol{\sigma}}/\delta)$$

and vectorized distributions are equivalent to multivariate normal (N) distributions

$$\mathrm{vec}(\mathcal{Y})|\mathrm{vec}(\mathcal{Z}_2) \sim N(\mathrm{vec}(\mathcal{M}) + (\boldsymbol{\Lambda}_K \otimes \cdots \otimes \boldsymbol{\Lambda}_1)\mathrm{vec}(\mathcal{Z}_2), \{\boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_1\}/\delta),$$

$$\mathrm{vec}(\mathcal{Z}_2) \sim N(\mathrm{vec}(0), \mathbf{D}^2_{\boldsymbol{\sigma}} \otimes \mathbf{I}_{d_1 \ldots d_{K-1}}/\delta).$$

Fix δ. Let $\mathrm{vec}(\mathcal{E})$ be a $\prod_{k=1}^K d_k$-dimensional vector following multivariate normal distribution,

$$\mathrm{vec}(\mathcal{E}) \sim N(\mathrm{vec}(\mathcal{M}), \{\boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_1\}/\delta),$$

Accordingly, we can build the $(2\prod_{k=1}^K d_k)$-dimensional joint multivariate normal distribution as follows,

$$\begin{pmatrix} \mathrm{vec}(\mathcal{E}) \\ \mathrm{vec}(\mathcal{Z}_2) \end{pmatrix} \Big| \delta \sim N\left(\begin{pmatrix} \mathrm{vec}(\mathcal{M}) \\ \mathrm{vec}(0) \end{pmatrix}, \frac{1}{\delta} \begin{pmatrix} \boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^2_{\boldsymbol{\sigma}} \otimes \mathbf{I}_{d_1 \ldots d_{K-1}} \end{pmatrix} \right).$$

We consider the transformation,

$$\begin{pmatrix} \text{vec}(\mathcal{Y}) \\ \text{vec}(\mathcal{Z}_2) \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{d_1 \dots d_K} & \{\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1\} \\ \mathbf{0} & \mathbf{I}_{d_1 \dots d_K} \end{pmatrix} \begin{pmatrix} \text{vec}(\mathcal{E}) \\ \text{vec}(\mathcal{Z}_2) \end{pmatrix}.$$

By properties of multivariate normal distribution,

$$\begin{pmatrix} \text{vec}(\mathcal{Y}) \\ \text{vec}(\mathcal{Z}_2) \end{pmatrix} \Big| \delta \sim N\Bigg(\begin{pmatrix} \text{vec}(\mathcal{M}) \\ \text{vec}(\mathit{0}) \end{pmatrix}, \\ \frac{1}{\delta} \left(\begin{array}{c|c} (\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \mathbf{\Lambda}_{K-1}^\intercal \mathbf{\Lambda}_{K-1} \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1) & \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1 \\ \hline \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1 & \mathbf{D}_{\boldsymbol{\sigma}}^2 \otimes \mathbf{I}_{d_1 \dots d_{K-1}} \end{array} \right) \Bigg).$$

Since $\delta \sim \text{Ga}(\nu/2, \nu/2)$, we can obtain the marginal distribution by integrating over δ as

$$\begin{pmatrix} \text{vec}(\mathcal{Y}) \\ \text{vec}(\mathcal{Z}_2) \end{pmatrix} \sim t\Bigg(\begin{pmatrix} \text{vec}(\mathcal{M}) \\ \text{vec}(\mathit{0}) \end{pmatrix}, \\ \left(\begin{array}{c|c} (\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \mathbf{\Lambda}_{K-1}^\intercal \mathbf{\Lambda}_{K-1} \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1) & \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1 \\ \hline \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1 & \mathbf{D}_{\boldsymbol{\sigma}}^2 \otimes \mathbf{I}_{d_1 \dots d_{K-1}} \end{array} \right), \nu \Bigg).$$

Our target pdf is $f(\text{vec}(\mathcal{Y}) | \text{vec}(\mathcal{Z}_2) > \mathit{0})$. Applying Bayes Theorem, we have

$$f(\text{vec}(\mathcal{Y}) | \text{vec}(\mathcal{Z}_2) > \mathit{0}) = \frac{f\text{vec}(\mathcal{Y}) P(\text{vec}(\mathcal{Z}_2) > \mathit{0} | \text{vec}(\mathcal{Y}))}{P(\text{vec}(\mathcal{Z}_2) > \mathit{0})}.$$

By properties of $(\prod_{k=1}^K d_k)$-variate-t distribution, we have marginal distributions.

$$\text{vec}(\mathcal{Y}) \sim t(\text{vec}(\mathcal{M}), (\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \mathbf{\Lambda}_{K-1}^\intercal \mathbf{\Lambda}_{K-1} \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1); \nu)$$

$$\text{vec}(\mathcal{Z}_2) \sim t(\text{vec}(\mathit{0}), \mathbf{D}_{\boldsymbol{\sigma}}^2 \otimes \mathbf{I}_{d_1 \dots d_{K-1}}; \nu).$$

and $P(\text{vec}(\mathcal{Z}_2) > \mathit{0}) = 2^{-(d_1 \dots d_K)}$, then target pdf is equal to

$$f(\text{vec}(\mathcal{Y}) | \text{vec}(\mathcal{Z}_2) > \mathit{0}) = 2^{(d_1 \dots d_K)} \\ \times t(\text{vec}(\mathcal{Y}) | \text{vec}(\mathcal{M}), (\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \mathbf{\Lambda}_{K-1}^\intercal \mathbf{\Lambda}_{K-1} \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1), \nu) \\ \times P(\text{vec}(\mathcal{Z}_2) > \mathit{0} | \text{vec}(\mathcal{Y})),$$

where $P(\text{vec}(\mathcal{Z}_2) > \mathit{0} | \text{vec}(\mathcal{Y}))$ can be derived as follows. From Lemma A.1 in [48], we have the conditional density of $f(\text{vec}(\mathcal{Z}_2) > \mathit{0} | \text{vec}(\mathcal{Y}))$ as $\text{vec}(\mathcal{Z}_2) > \mathit{0} | \text{vec}(\mathcal{Y}) \sim t(\mathbf{m}_{\mathcal{Z}_2|\mathcal{Y}}, \mathbf{R}_{\mathcal{Z}_2|\mathcal{Y}}, \nu_{\mathcal{Z}_2|\mathcal{Y}})$.

$$\mathbf{m}_{\mathcal{Z}_2|\mathcal{Y}} = [(\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1)\{(\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1}] \text{vec}(\mathcal{Y})^* \\ \mathbf{R}_{\mathcal{Z}_2|\mathcal{Y}} = \frac{\nu + \text{vec}(\mathcal{Y})^{*\intercal} \{(\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1} \text{vec}(\mathcal{Y})^*}{\nu + \prod_{k=1}^K d_k} \\ \times \mathbf{D}_{\boldsymbol{\sigma}}^2 \otimes \mathbf{I}_{d_1 \dots d_{K-1}} - (\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1)\{(\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1} \\ \times (\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1), \quad \text{vec}(\mathcal{Y})^* = \text{vec}(\mathcal{Y}) - \text{vec}(\mathcal{M}), \quad \nu_{\mathcal{Z}_2|\mathcal{Y}} = \nu + \prod_{k=1}^K d_k,$$

which provides $P(\text{vec}(\mathcal{Z}_2) > 0 | \text{vec}(\mathcal{Y})) = T(\mathbf{z}_2^* | \Phi, \nu + \prod_{k=1}^K d_k)$, where

$$\mathbf{z}_2^* = \mathbf{m}_{\mathcal{Z}_2|\mathcal{Y}} \times$$
$$\sqrt{\frac{\nu + \prod_{k=1}^K d_k}{\nu + (\text{vec}(\mathcal{Y}) - \text{vec}(\mathcal{M}))^\intercal \{(\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1} (\text{vec}(\mathcal{Y}) - \text{vec}(\mathcal{M}))}},$$
$$\text{and } \Phi = \mathbf{D}_{\boldsymbol{\sigma}}^2 \otimes \mathbf{I}_{d_1 \ldots d_{K-1}} - (\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1)\{(\mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1) + (\mathbf{\Lambda}_K^\intercal \mathbf{D}_\sigma^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1^\intercal \mathbf{\Lambda}_1)\}^{-1}$$
$$\times (\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1)$$

Thus, the proof is complete. □

B.2 Posterior Distributions for GAAD Study

B.2.1 Posterior Distributions for BSTT

1. We sample vectorized tensor coefficients from multivariate normal distribution. Let $\boldsymbol{\beta} = \text{vec}(\mathcal{B})$.

$$p(\boldsymbol{\beta}|-) \sim N_{TSJp}\left(\text{A}^{-1}\left\{\text{vec}(S_{xy}) - \text{vec}(S_{xw})\right\}, \text{A}^{-1}\right), \text{ where}$$
$$S_{xy} = \sum_{i=1}^n \mathbf{x}_i \mathbf{y}_i^\intercal \mathbf{\Sigma}^{-1}, \; S_{xw} = \sum_{i=1}^n \mathbf{x}_i \mathbf{z}_{2i}^\intercal (\mathbf{\Lambda}_K \otimes \mathbf{R}_2^{1/2} \otimes \mathbf{R}_1^{1/2}) \mathbf{\Sigma}^{-1},$$
$$\text{A} = \left(S_{xx} \otimes \mathbf{\Sigma}^{-1} + \frac{1}{\text{c}^4} \mathbf{I}_{TSJp}\right), \; \mathbf{\Sigma} = (\mathbf{D}_\sigma \mathbf{R}_3 \mathbf{D}_\sigma \otimes \mathbf{R}_2 \otimes \mathbf{R}_1)/\delta_i, \; S_{xx} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\intercal,$$

and $\mathbf{B}_{(4)}$ is matricization along the fourth mode of $\mathcal{B}$.

2. For each patient i, we sample vectorized latent variable $\mathbf{z}_{2i}$ from truncated multivariate-t distribution. Define $\mathbf{z}_{2i} = \text{vec}(\mathcal{Z}_{2i})$.

$$p(\mathbf{z}_{2i}|-) \sim Trt_{TSJ}\left(\mathbf{m}_{\mathbf{z}_{2i}}, \mathbf{\Sigma}_{\mathbf{z}_{2i}}, \nu_{\mathbf{z}_{2i}}\right) I(\mathbf{z}_{2i} > 0), \text{ where}$$
$$\mathbf{m}_{\mathbf{z}_{2i}} = \left\{\mathbf{\Lambda}_K \left(\mathbf{D}_\sigma \mathbf{R}_3 \mathbf{D}_\sigma/\delta_i + \mathbf{\Lambda}_K^2\right)^{-1} \otimes \mathbf{R}_2^{-1/2} \otimes \mathbf{R}_1^{-1/2}\right\} \left(\text{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i\right), \text{ and } \mathbf{\Lambda}_K = \text{diag}(\lambda_1, \lambda_2)$$
$$\mathbf{\Sigma}_{\mathbf{z}_{2i}} = \frac{\left(\text{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i\right)^\intercal \left\{\left(\mathbf{D}_\sigma \mathbf{R}_3 \mathbf{D}_\sigma/\delta_i + \mathbf{\Lambda}_K^2\right)^{-1} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\right\} \left(\text{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i\right) + \nu}{TSJ + \nu}$$
$$\times (\mathbf{I}_{TSJ} + \mathbf{\Lambda}_K (\mathbf{D}_\sigma \mathbf{R}_3 \mathbf{D}_\sigma/\delta_i)^{-1} \mathbf{\Lambda}_K \otimes \mathbf{I}_{TS})^{-1}$$

$\nu_{\mathbf{z}_{2i}} = TSJ + \nu$, where "$Trt_{TSJ}$" denotes truncated TSJ-variate-t distribution.

3. We sample latent variable δ_i from gamma distribution for $i = 1, \ldots, n$.

$$p(\delta_i|-) \sim Ga(a_\delta, b_{i,\delta}), \text{ where}$$
$$a_\delta = \frac{\nu}{2} + TSJ$$
$$b_{i,\delta} = \frac{\nu}{2} + \frac{1}{2}\text{tr}\Bigg[\left\{\text{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i - \left(\mathbf{\Lambda}_K \otimes \mathbf{R}_2^{1/2} \otimes \mathbf{R}_1^{1/2}\right)\mathbf{z}_{2i}\right\}^\intercal$$
$$\times (\mathbf{D}_\sigma \mathbf{R}_3 \mathbf{D}_\sigma)^{-1} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1} \left\{\text{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i - \left(\mathbf{\Lambda}_K \otimes \mathbf{R}_2^{1/2} \otimes \mathbf{R}_1^{1/2}\right)\mathbf{z}_{2i}\right\} \Bigg]$$

4. We sample λ_j for $j = 1, 2$ (1 for PPD and 2 for CAL) from univariate normal distribution.

$$p(\lambda_j) \sim N\left(\frac{G_\gamma}{H_{\gamma,\gamma}}, \frac{1}{H_{\gamma,\gamma}}\right),$$
$$\text{where } H_{\gamma,\gamma} = \sum_{i=1}^{n} \left[(|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J)^\intercal \mathbf{\Sigma}^{-1} (|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J) + \mathbf{I}_{J^2}\right]_{j^2,j^2}, \text{ and}$$
$$G_\gamma = \sum_{i=1}^{n} \Big\{ \left[(\mathbf{y}_i^\intercal - \mathbf{x}_i \mathbf{B}_{(4)}) \mathbf{\Sigma}^{-1} (|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J) + \mathbf{1}_0\right]_{j^2}$$
$$- \sum_{j^2 \neq j'^2} \left[(|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J)^\intercal \mathbf{\Sigma}^{-1} (|\mathbf{Z}_{2i(3)}|^\intercal \otimes \mathbf{I}_J) + \mathbf{I}_{J^2}\right]_{j^2,j'^2} \lambda_{j'} \Big\},$$
$$\text{where } (j^2, j^2) \text{ represents } j^2\text{-th diagonal elements of } J^2 \times J^2 \text{ matrix.}$$

5. We sample ρ_1, ρ_2, and ρ_3 using MH algorithm. Let $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$, and $\boldsymbol{\rho}^* = (\rho_1^*, \rho_2^*, \rho_3^*)$. We apply MH algorithm to update ρ_1, ρ_2, and ρ_3 jointly with proposal densities as ρ_1^*, ρ_2^* $\rho_3^* \sim Beta(2, 2)$.

Set the initial values $\rho_1^{(1)}$, $\rho_2^{(1)}$ and $\rho_3^{(1)}$. For each $s = 2, \ldots, S$,

a) sample $\rho_1^{new} \sim q(\cdot|\rho_1^{(s-1)})$, $\rho_2^{new} \sim g(\cdot|\rho_2^{(s-1)})$, and $\rho_3^{new} \sim g(\cdot|\rho_3^{(s-1)})$.

b) set $\rho_1^{(s)} = \rho_1^{new}$ with the probability $\min\left\{1, \alpha_1\left(\rho_1^{(s-1)}, \rho_1^{new}\right)\right\}$,

where $\alpha_1(\rho_1^{(s-1)}, \rho_1^{new}) = \dfrac{p(\rho_1^{new}|-)q(\rho_1^{(s-1)}|\rho_1^{new})}{p(\rho_1^{(s-1)}|-)q(\rho_1^{new}|\rho_1^{(s-1)})}$,

otherwise set $\rho_1^{(s)} = \rho_1^{(s-1)}$.

c) set $\rho_2^{(s)} = \rho_2^{new}$ with the probability $\min\left\{1, \alpha_2\left(\rho_2^{(s-1)}, \rho_2^{new}\right)\right\}$,

where $\alpha_2(\rho_2^{(s-1)}, \rho_2^{new}) = \dfrac{p(\rho_2^{new}|-)q(\rho_2^{(s-1)}|\rho_2^{new})}{p(\rho_2^{(s-1)}|-)q(\rho_2^{new}|\rho_2^{(s-1)})}$,

otherwise set $\rho_2^{(s)} = \rho_2^{(s-1)}$.

d) set $\rho_3^{(s)} = \rho_3^{new}$ with the probability $\min\left\{1, \alpha_3\left(\rho_3^{(s-1)}, \rho_3^{new}\right)\right\}$,

where $\alpha_3(\rho_3^{(s-1)}, \rho_3^{new}) = \dfrac{p(\rho_3^{new}|-)q(\rho_3^{(s-1)}|\rho_3^{new})}{p(\rho_3^{(s-1)}|-)q(\rho_3^{new}|\rho_3^{(s-1)})}$,

otherwise set $\rho_3^{(s)} = \rho_3^{(s-1)}$.

6-1. We sample $\frac{1}{\sigma_1^2}$ (scale parameter for PPD response) from gamma distribution.

$$p\left(\frac{1}{\sigma_1^2}\middle|-\right) \sim Ga(nTS + g_1, \phi_1), \text{ where}$$

$$\phi_1 = \frac{1}{2}\sum_{i=1}^{n}\left\{\mathbf{S}^\intercal\left[\delta_i\begin{pmatrix} \frac{1}{(1-\rho_3)(1+\rho_3)} & -\frac{\sigma_1\rho_3}{\sigma_2(1-\rho_3)(1+\rho_3)} \\ -\frac{\sigma_1\rho_3}{\sigma_2(1-\rho_3)(1+\rho_3)} & \frac{\sigma_1^2}{\sigma_2^2(1-\rho_3)(1+\rho_3)}\end{pmatrix} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\right]\mathbf{S}\right.$$
$$\left. + \delta_i \text{vec}(|\mathcal{Z}_{2i}|)^\intercal \text{diag}(1, \sigma_1^2/\sigma_2^2, \ldots, 1, \sigma_1^2/\sigma_2^2)\text{vec}(|\mathcal{Z}_{2i}|)\right\} + g_2,$$

and $\mathbf{S} = \text{vec}(\mathcal{Y}_i) - \mathbf{B}_{(4)}^\intercal \mathbf{x}_i - \left(\mathbf{\Lambda}_K \otimes \mathbf{R}_2^{1/2} \otimes \mathbf{R}_1^{1/2}\right)\text{vec}(|\mathcal{Z}_{2i}|)$.

6-2. We sample $\frac{1}{\sigma_2^2}$ (scale parameter for CAL response) from gamma distribution.

$$p\left(\frac{1}{\sigma_2^2}\middle|-\right) \sim Ga(nTS + g_2, \phi_2), \text{ where}$$

$$\phi_2 = \frac{1}{2}\sum_{i=1}^{n}\left\{\mathbf{S}^\intercal\left[\delta_i\begin{pmatrix} \frac{\sigma_2^2}{\sigma_1^2(1-\rho_3)(1+\rho_3)} & -\frac{\sigma_2\rho_3}{\sigma_1(1-\rho_3)(1+\rho_3)} \\ -\frac{\sigma_2\rho_3}{\sigma_1(1-\rho_3)(1+\rho_3)} & \frac{1}{(1-\rho_3)(1+\rho_3)}\end{pmatrix} \otimes \mathbf{R}_2^{-1} \otimes \mathbf{R}_1^{-1}\right]\mathbf{S}\right.$$
$$\left. + \delta_i \text{vec}(|\mathcal{Z}_{2i}|)^\intercal \text{diag}(\sigma_2^2/\sigma_1^2, 1, \ldots, \sigma_2^2/\sigma_1^2, 1)\text{vec}(|\mathcal{Z}_{2i}|)\right\} + g_2.$$

7. We sample ν using MH algorithm. The conditional distribution is

$$p(\nu|-) \propto \frac{(\frac{\nu}{2})^{\frac{n\nu}{2}}}{[\Gamma(\nu/2)]^n}\left(\prod_{i=1}^{n}\delta_i^{\frac{\nu}{2}-1}\right)\exp\left(-\frac{\nu}{2}\sum_{i=1}^{n}\delta_i\right)\frac{1}{\sigma_1\sigma_2}\sqrt{\psi'(\frac{\nu}{2}) - \psi'(\frac{\nu+1}{2}) - \frac{2(\nu+3)}{\nu(\nu+1)^2}}.$$

We follow the suggestion by [30], the proposal distribution is truncated normal distribution with the support of (0, 300), $N_{(0<\nu<300)}(a_\nu, b_\nu^2)$, where $a_\nu = \nu - \frac{g'(\nu)}{g''(\nu)}$, $b_\nu^2 = -\frac{1}{g''(\nu)}$, and $g(\nu) = \log(p(\nu|-))$.

B.3 Additional Results of GAAD Data Analysis

B.3.1 Results with TN Prior

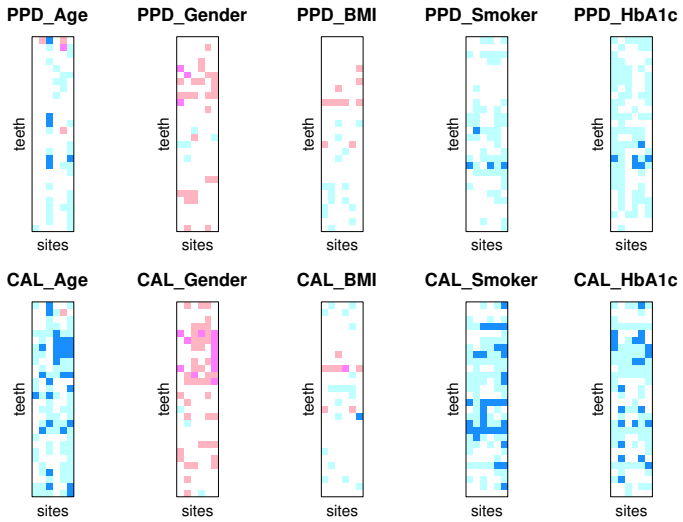

Figure B.1: Fitting the BTN Model with Tensor Normal prior for $\mathcal{B}$ to the GAAD data. Plotted are the D-statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations. Strong (moderate) evidence of the positive effects are shaded in dark blue (light blue), neutral evidence as white. Strong (moderate) evidence of the negative effects are shaded in dark pink (light pink).

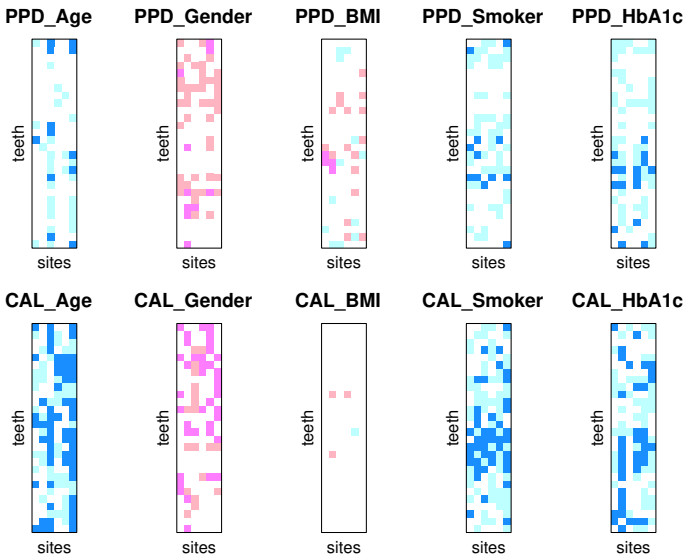

Figure B.2: Fitting the BTT Model with Tensor Normal prior for $\mathcal{B}$ to the GAAD data. Plotted are the D-statistics heatmaps of the covariate associations on PPD (top row plates) and CAL (bottom row plates), for the various teeth (28 rows in each plate) and site (6 columns in each plate) combinations.

B.3.2 Results with TSSL Prior

Table B.1: Bayesian analysis of GAAD using the tensor normal model and tensor-t model with TSSL prior of (5.3) in the main text: Posterior summaries of the overall effects of each covariate on PPD and CAL.

BTN	PPD	Median	SD	CI	BTT	PPD	Median	SD	CI
	Intercept	1.7604	0.3445	(1.0989, 2.4187)		Intercept	1.8023	0.2160	(1.4036, 2.2266)
	Age	0.0018	0.0054	(-0.0087, 0.0121)		Age	0.0011	0.0034	(-0.0057, 0.0077)
	Gender	-0.0908	0.1371	(-0.3674, 0.1645)		Gender	-0.1259	0.0847	(-0.2950, 0.0403)
	BMI	-0.0083	0.1153	(-0.2114, 0.2427)		BMI	-0.0001	0.0767	(-0.1432, 0.1437)
	Smoker	0.1664	0.1234	(-0.0759, 0.3971)		Smoker	0.1597	0.0761	(0.0145, 0.3015)
	HbA1c	0.2194	0.1315	(-0.0473, 0.4768)		HbA1c	0.2335	0.0776	(0.0863, 0.3822)
	CAL	Median	SD	CI		CAL	Median	SD	CI
	Intercept	1.2721	0.2559	(0.7443, 1.7239)		Intercept	1.2574	0.1978	(0.8741, 1.6530)
	Age	0.0095	0.0041	(0.0023, 0.0181)		Age	0.0097	0.0032	(0.0035, 0.0158)
	Gender	-0.2172	0.1046	(-0.4178, -0.0062)		Gender	-0.2125	0.0764	(-0.3356, -0.0602)
	BMI	0.0213	0.0901	(-0.1504, 0.1879)		BMI	0.0130	0.0693	(-0.1319, 0.1330)
	Smoker	0.3275	0.0913	(0.1540, 0.5079)		Smoker	0.3281	0.0737	(0.1937, 0.4748)
	HbA1c	0.3086	0.0967	(0.1149, 0.4891)		HbA1c	0.3018	0.0747	(0.1520, 0.4482)

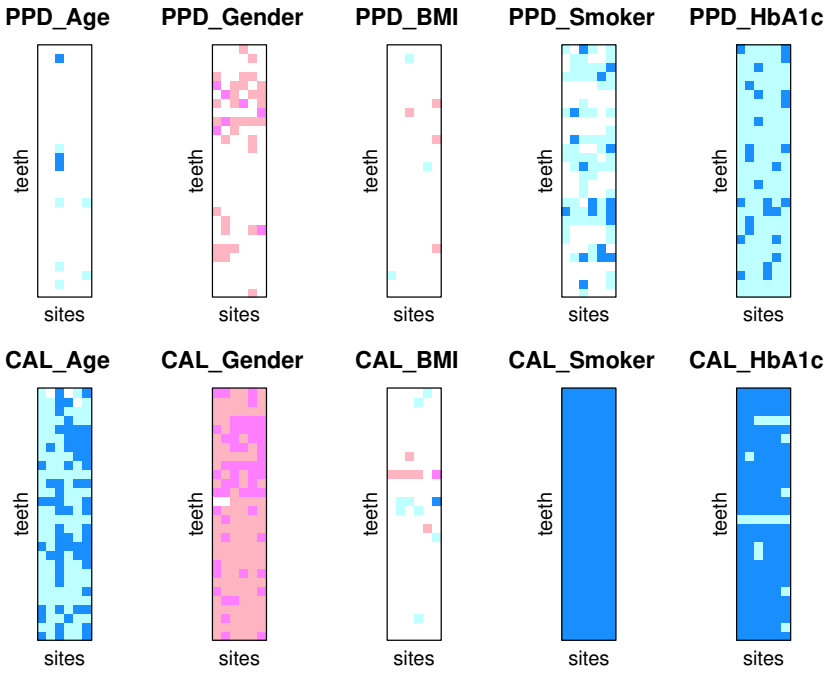

Figure B.3: Analysis of GAAD using BTN model with TSSL prior for $\mathcal{B}$: Heatmap of D-statistics of the covariate effects (columns of plates) on PPD (top row of plates) and CAL (bottom row of plates) on various teeth (6 rows in each plate) and sites (28 columns in each plate) combinations.

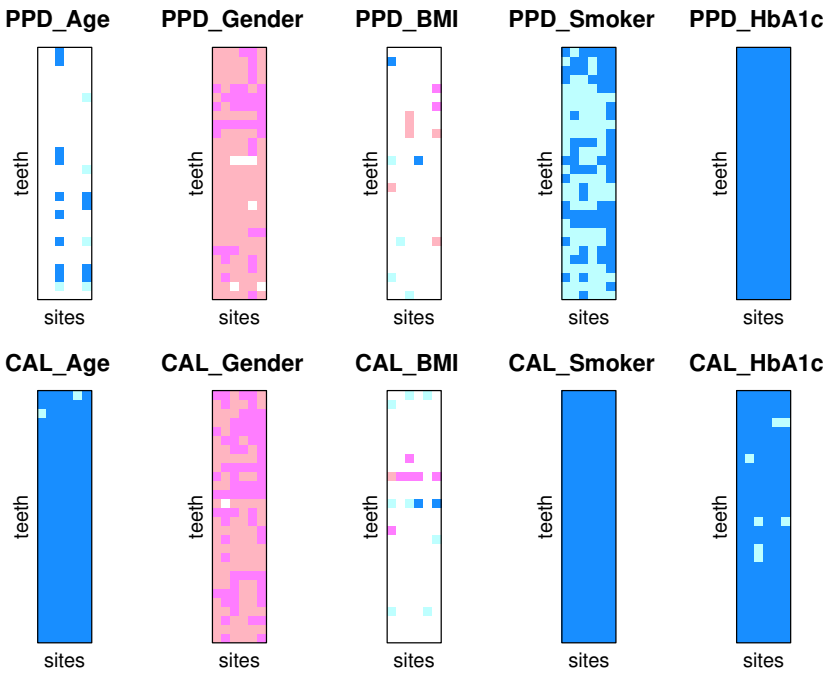

Figure B.4: Analysis of GAAD using BTT model with TSSL prior for $\mathcal{B}$: Heatmap of D-statistics of the covariate effects (columns of plates) on PPD (top row of plates) and CAL (bottom row of plates) on various teeth (6 rows in each plate) and sites (28 columns in each plate) combinations.

REFERENCES

[1] Adelchi Azzalini and Antonella Capitanio. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602, 1999.

[2] Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):367–389, 2003.

[3] Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

[4] Dipankar Bandyopadhyay, Victor H Lachos, Carlos A Abanto-Valle, and Pulak Ghosh. Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. *Statistics in Medicine*, 29(25):2643–2655, 2010.

[5] Dipankar Bandyopadhyay, Victor H Lachos, Luis M Castro, and Dipak K Dey. Skew-normal/independent linear mixed models for censored responses with applications to hiv viral loads. *Biometrical Journal*, 54(3):405–425, 2012.

[6] Peter J Basser and Sinisa Pajevic. A normal distribution for tensor-valued random variables: Applications to diffusion tensor MRI. *IEEE Transactions on Medical Imaging*, 22(7):785–794, 2003.

[7] Apurva Bhingare, Debajyoti Sinha, Debdeep Pati, Dipankar Bandyopadhyay, and Stuart R Lipsitz. Semiparametric bayesian latent variable regression for skewed multivariate data. *Biometrics*, 75(2):528–538, 2019.

[8] Márcia D Branco and Dipak K Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79(1):99–113, 2001.

[9] Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

[10] Shu-Ching Chang and Dale L Zimmerman. Skew-normal antedependence models for skewed longitudinal data. *Biometrika*, 103(2):363–376, 2016.

[11] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[12] International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes care*, 32(7):1327–1334, 2009.

[13] R Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010.

[14] ML Darby and MM Walsh. *Dental Hygiene: Theory and Practice (1st edn)*. W. B. Saunders Company: U.S.A., 1995.

[15] Arthur Pentland Dempster. *Elements of continuous multivariate analysis*. Reading, Mass. : Addison-Wesley Pub. Co., 1969.

[16] Morris L Eaton. On the projections of isotropic distributions. *The Annals of Statistics*, 9(2):391–400, 1981.

[17] Robert Engle and Bryan Kelly. Dynamic equicorrelation. *Journal of Business & Economic Statistics*, 30(2):212–228, 2012.

[18] K-T Fang, S Kotz, and K-W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London, 1990.

[19] Jyotika K Fernandes, Ryan E Wiegand, Carlos F Salinas, Sara G Grossi, John J Sanders, Maria F Lopes-Virella, and Elizabeth H Slate. Periodontal disease status in gullah african americans with type 2 diabetes living in south carolina. *Journal of periodontology*, 80(7):1062–1068, 2009.

[20] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[21] Sharmistha Guha and Rajarshi Guhaniyogi. Bayesian generalized sparse symmetric tensor-on-vector regression. *Technometrics*, pages 1–11, 2020.

[22] Rajarshi Guhaniyogi, Shaan Qamar, and David B Dunson. Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1):2733–2763, 2017.

[23] Rajarshi Guhaniyogi and Daniel Spencer. Bayesian tensor response regression with an application to brain activation studies. Technical report, Technical report, UCSC. 2, 13, 2018.

[24] Arjun K Gupta, Tamas Varga, and Taras Bodnar. *Elliptically contoured models in statistics and portfolio theory*. Springer, 2013.

[25] Daojiang He, Dongchu Sun, and Lei He. Objective bayesian analysis for the student-t linear regression. *Bayesian Analysis*, 2020.

[26] Peter D Hoff. Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196, 2011.

[27] Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169, 2015.

[28] J. G Ibrahim, M.-H Chen, and D. Sinha. *Bayesian Survival Analysis.* New York: Springer-Verlag, 2001.

[29] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[30] Shuaimin Kang, Guangying Liu, Howard Qi, and Min Wang. Bayesian variance changepoint detection in linear models with symmetric heavy-tailed errors. *Computational Economics*, 52(2):459–477, 2018.

[31] Douglas Kelker. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 419–430, 1970.

[32] Tamara G Kolda and Brett W Bader. Tensor decompositions and Applications. *SIAM review*, 51(3):455–500, 2009.

[33] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

[34] Xiaoshan Li, Da Xu, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545, 2018.

[35] John C Liechty, Merrill W Liechty, and Peter Müller. Bayesian correlation estimation. *Biometrika*, 91(1):1–14, 2004.

[36] Antonio R Linero and Michael J Daniels. Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science: A review journal of the Institute of Mathematical Statistics*, 33(2):198–213, 2018.

[37] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

[38] Yanyuan Ma, Marc G Genton, and Anastasios A Tsiatis. Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *Journal of the American Statistical Association*, 100(471):980–989, 2005.

[39] Robb J Muirhead. Aspects of multivariate statistical analysis. *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1982, 656*, 1982.

[40] Carlos Antonio Negrato and Olinda Tarzia. Buccal alterations in diabetes mellitus. *Diabetology & metabolic syndrome*, 2(1):1–11, 2010.

[41] Roy C Page and Paul I Eke. Case definitions for use in population-based surveillance of periodontitis. *Journal of periodontology*, 78:1387–1399, 2007.

[42] Georgia Papadogeorgou, Zhengwu Zhang, and David B Dunson. Soft tensor regression. *arXiv preprint arXiv:1910.09699*, 2019.

[43] Debdeep Pati, Anirban Bhattacharya, Natesh S Pillai, David Dunson, et al. Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102–1130, 2014.

[44] Karl Pearson. X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, (186):343–414, 1895.

[45] Brian J Reich and Dipankar Bandyopadhyay. A latent factor model for spatial data with informative missingness. *The Annals of Applied Statistics*, 4(1):439–459, 2010.

[46] Brian J Reich, Dipankar Bandyopadhyay, and Howard D Bondell. A nonparametric spatial model for periodontal data with nonrandom missingness. *Journal of the American Statistical Association*, 108(503):820–831, 2013.

[47] Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

[48] Sujit K Sahu, Dipak K Dey, and Márcia D Branco. A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150, 2003.

[49] Daniel Spencer, Rajarshi Guhaniyogi, and Raquel Prado. Bayesian mixed effect sparse tensor response regression model with joint estimation of activation and connectivity. *arXiv preprint arXiv:1904.00148*, 2019.

[50] Will Wei Sun and Lexin Li. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

[51] Panagiotis Symeonidis and Andreas Zioupos. *Matrix and Tensor Factorization Techniques for Recommender Systems*, volume 1. Springer, 2016.

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

[53] Min Wang and Mingan Yang. Posterior property of student-t linear regression model using objective priors. *Statistics & Probability Letters*, 113:23–29, 2016.

[54] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11:3571–3594, 2010.

[55] Lin Zhang and Dipankar Bandyopadhyay. A graphical model for skewed matrix-variate non-randomly missing data. *Biostatistics*, 21(2):e80–e97, 2020.

[56] Xin Zhang and Lexin Li. Tensor envelope partial least-squares regression. *Technometrics*, 59(4):426–436, 2017.

[57] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

BIOGRAPHICAL SKETCH

The author obtained a Bachelor of Economics degree in Statistics from Sungkyunkwan University, Seoul, South Korea, in 2016. He then pursued his Ph.D. degree at the Department of Statistics, Florida State University.