FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

BAYESIAN ADDITIVE REGRESSION TREES FOR MULTIVARIATE RESPONSES

By

SEUNGHA UM

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

2021

Seungha Um defended this dissertation on July 6, 2021.
The members of the supervisory committee were:

Debajyoti Sinha
Professor Directing Dissertation

Richard Bertram
University Representative

Antonio Linero
Committee Member

Eric Chicken
Committee Member

Andres Felipe Barrientos
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 dearest husband Hwiyoung, my beloved child Hazel, and my best friend Latte, whose love and support mean the world to me.

ACKNOWLEDGMENTS

I would first like to express my deepest appreciation to my advisors Dr. Debajyoti Sinha and Dr. Antonio Linero for their academic guidance, unwavering support and enthusiastic encouragement. Besides my advisors, I would like to thank Dr. Dipankar Bandyopadhyay for his collaborations and suggestions. It has been a great honor to work with them. I would also like to express my sincere gratitude to my doctoral committee members, Dr. Eric Chicken, Dr. Andres Felipe Barrientos, and Dr. Richard Bertram, each of whom provided constructive comments and encouragement, which helped me in enriching and improving this manuscript.

I must also express my profound gratitude to my family who supported me through years of graduate school and provided me with constant support. I sincerely could not have made it through without their undying faith in my ability and all you have done for me to get here. My little princess Hazel inspires me to be the greatest version of myself. She is my reason for life and there is no limit to my love for her.

Last but not least, I would like to thank my husband Hwiyoung for his encouragement and unfailing support. He is my best friend, my greatest support, my biggest comfort and strongest motivation. My husband made this happen. There are no words to convey how important you are to me and how much I appreciate and love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation introduces a nonparametric regression approach for multivariate skewed responses using Bayesian additive regression trees (BART). Existing BART methods use ensembles of decision trees to model a mean function, and have become popular recently due to their high prediction accuracy and ease of use. The usual assumption of a univariate Gaussian error distribution, however, is restrictive in many biomedical applications. Motivated by an oral health study, we provide a useful extension of BART, called the `skewBART` model, to address this problem. We then extend `skewBART` to allow for multivariate responses, with information shared across the decision trees associated with different responses within the same subject. The methodology accommodates within subject association and allows different skewness parameters for different responses. We illustrate the benefit of our multivariate `skewBART` model by analyzing the oral health study with highly skewed multiple responses.

Secondly, BART model for analyzing mixed correlated continuous and count responses is proposed in this dissertation. Although Bayesian approaches are favored to model the uncertainty of estimation and to incorporate prior information, the limitation of developing simple and efficient algorithms for posterior computation has made Bayesian analysis of counts underdeveloped. However, recent latent variable construction based on Polya-Gamma random variables makes Bayesian inference appealing in the negative-binomial model with efficient sampling derived by exploiting conditional conjugacy. By utilizing Pólya-Gamma latent variable construction, the proposed BART model accounts for associations between count and continuous responses. Moreover, an efficient gibbs sampling is derived by exploiting conditional conjugacy since the negative binomial likelihood can be represented as a mixture of normals with Polya-Gamma mixing distribution. The utility of the models is demonstrated by analyzing the oral health study with mixed multiple responses.

CHAPTER 1

INTRODUCTION

1.1 Motivation

In many biomedical and clinical studies, multivariate skewed responses are commonly observed. For the multivariate responses, it is often of interest to identify unknown set of features associated with the responses. In our motivating example of periodontal disease (PD) status, three of the most popular markers are (a) the mean clinical attachment loss (CAL), (b) the mean periodontal pocket depth (PPD) and (c) the number of sites bleeding on probing (BOP) for each subject. The mean PPD and the mean CAL are continuous variables and highly skewed. CAL is assessed by the distance from a fixed reference point to the base of the pocket, and PPD is the distance from the gingival margin to the bottom of the gingival pocket. BOP is an indicator of tissue whether a particular site bled with the application of a dental probe. All three markers are usually measured at six pre-specified sites for each tooth. They are highly skewed both marginally and jointly.

The mean CAL and the mean PPD are continuous variables and highly skewed with the skewness of CAL being different from the skewness of PPD. The BOP is a count variable which is associated with two other markers. The goal of our motivating example is to quantify the disease status, and to study the associations between disease status and patient-level covariates such as age, BMI, gender, HbAlC and smoking status. Also, it is crucial to jointly model CAL, PPD and BOP simultaneously in terms of a latent periodontal health factor. The joint model is required to account for associations between count and continuous responses.

1.2 Literature review

1.2.1 Bayesian additive regression trees model

The BART framework introduced by Chipman et al.[12] models $f(\boldsymbol{x}_i)$ in (2.1) as a sum of m trees

$$f(\boldsymbol{x}_i) = \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) \ , \tag{1.1}$$

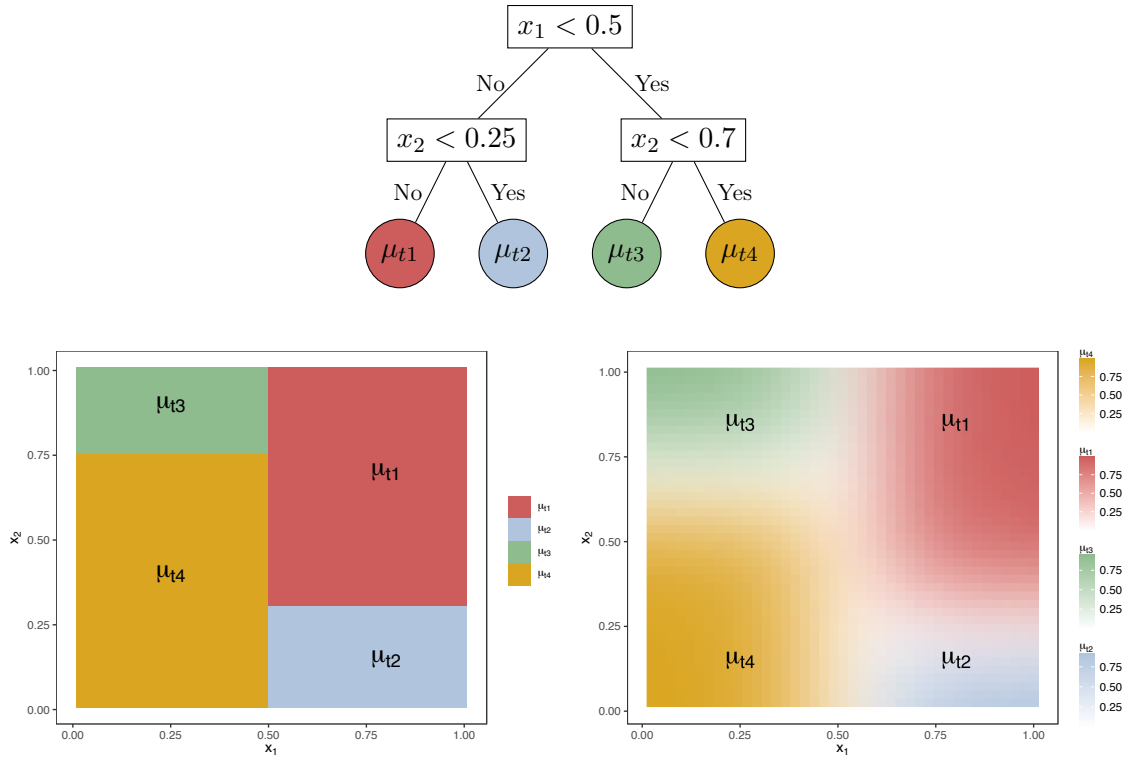

Figure 1.1: Top: an example of a binary tree $\mathcal{T}_t$ where the terminal nodes are labeled with the corresponding parameters $\mu_{t\ell}$. Bottom: the corresponding partition of the predictor space $\mathcal{X} = [0,1]^2$ from BART (left panel) and SBART (right panel) when $\kappa = 0.1$.

where $\mathcal{T}_t$ is a binary tree structure with n_t leaf nodes and $\mathcal{M}_t = \{\mu_{t1}, \ldots, \mu_{tn_t}\}$ is a collection of leaf parameters for $\mathcal{T}_t$ with h_ϵ being a Gaussian density. The function $g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t)$ returns $\sum_{\ell=1}^{n_t} \mu_{t\ell}\phi(x_i; \mathcal{T}_t, \ell)$ where $\phi(x_i; \mathcal{T}_t, \ell)$ is the indicator that x_i is associated with leaf node ℓ in $\mathcal{T}_t$, i.e., $g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) = \mu_{t\ell}$ if-and-only-if $\boldsymbol{x}$ is associated to leaf ℓ of tree t; for example, if $\boldsymbol{x} = (0.62, 0.45)$ in Figure 1.1 (bottom-left), then $g(\boldsymbol{x}, \mathcal{T}_t, \mathcal{M}_t) = \mu_{t1}$. The tree $\mathcal{T}_t$ consists of interior splitting rules of the form $[x_j \leq C_b]$, which induces a partition on the covariate space. At each interior node, the splitting rule uses predictor j with probability s_j, where $s = (s_1, \ldots, s_p)$ is a probability vector. The coordinate C_b of the splitting rule is assigned a uniform prior over its possible values.

To avoid overfitting, each tree is given a *regularization prior* which shrinks the individual leaf parameters to 0 so that each tree contributes only a small portion of the overall fit. It is regularized by the depth of each tree and the prior distribution of leaf parameters. The probability that each node at depth d is non-terminal is given by $\alpha(1+d)^{-\beta}$ for hyperparameters $\alpha > 0$ and $\beta > 0$ (Figure

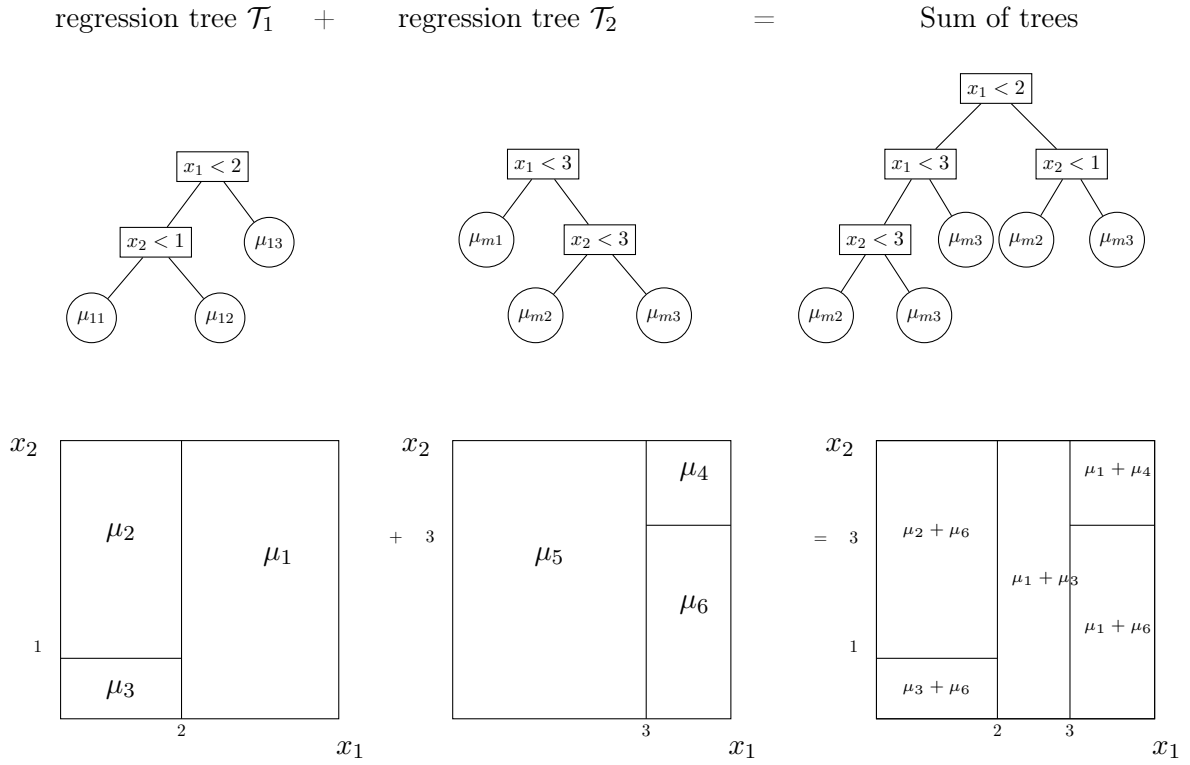

Figure 1.2: Top: An example of sum of two binary trees where terminal nodes are labeled with the corresponding parameters. Bottom: The corresponding partition of the predictor from BART.

1.2.1). The default prior specification[12] sets $\alpha = 0.95$ and $\beta = 2$, which encourages the trees to be shallow (rarely exceeding depth 2–3). The leaf parameters $\mu_{t\ell}$ are given iid $\mathcal{N}(0, \sigma_\mu^2/m)$ priors, ensuring that the prior variance of $f(\boldsymbol{x})$ is constant as the number of trees increases.

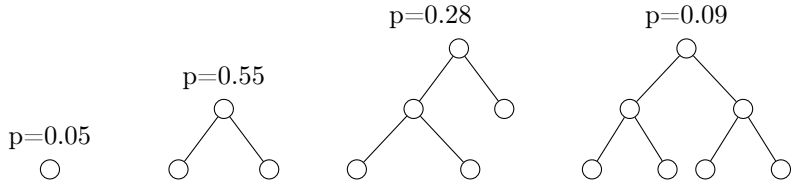

Figure 1.3: An example of tree depth with default prior specification, $\alpha = 0.95$, $\beta = 2$.

To define a BART model (1.1) for the non-parametric response regression in (2.1), we must specify a likelihood for the data $(\mathbf{Y}, \mathbf{X})$, a prior on the unknown model parameters related to the tree structures, and a prior on the parameters γ of the unknown error distribution $h_\epsilon(\cdot \mid \gamma)$. Bayesian computation then proceeds by using Markov chain Monte Carlo (MCMC) techniques to sample from

the joint posterior $p(\mathcal{T}_1, \mathcal{M}_1, \ldots, \mathcal{T}_m, \mathcal{M}_m, \sigma^2 \mid \mathbf{Y}, \mathbf{X})$, which is proportional to the likelihood times the joint prior $\pi(\theta)$ of all the unknown parameters θ. When the error distribution is $\mathcal{N}(0, \sigma^2)$, the prior consists of a conjugate inverse gamma prior $\pi(\sigma^2)$ for σ^2, independent priors $\pi(\mathcal{T}_t)$ for the tree structures $\mathcal{T}_t$, and conditionally-independent priors for the terminal node parameters. For the leaf node set $\mathcal{M}_t = \{\mu_{t\ell} : 1 \geq \ell \geq n_t\}$ conditional on the tree structure $\mathcal{T}_t$, a conjugate normal distribution $\mathcal{N}(0, \sigma_\mu^2/m)$ is widely used as the common independent prior $\pi(\mu_{tl})$. The resulting joint prior distribution factors as

$$\begin{aligned}
\pi((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \sigma^2) &= \prod_{t=1}^{m} \Big[\pi(\mathcal{T}_t, \mathcal{M}_t)\Big] \pi(\sigma^2) \\
&= \prod_{t=1}^{m} \Big[\pi(\mathcal{M}_t|\mathcal{T}_t)\pi(\mathcal{T}_t)\Big] \pi(\sigma^2) \qquad (1.2) \\
&= \prod_{t=1}^{m} \prod_{\ell=1}^{n_t} \Big[\pi(\mu_{t\ell}|\mathcal{T}_t)\pi(\mathcal{T}_t)\Big] \pi(\sigma^2).
\end{aligned}$$

The marginal likelihood (conditional posterior distribution) of $\mathcal{T}_t$ given $\mathcal{T}_{-t} = \{\mathcal{T}_k : k \neq t\}$, and similarly defined $\mathcal{M}_{-t}$ is obtained as

$$L(\mathcal{T}_t \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}) = \int \prod_{i=1}^{n} p(Y_i|\mathcal{T}_1, \mathcal{M}_1, \ldots, \mathcal{T}_m, \mathcal{M}_m, \sigma^2) p(\mathcal{M}_t|\mathcal{T}_t, \sigma^2) d\mathcal{M}_t \qquad (1.3)$$

which fortunately has a closed form expression in the regression setting. This allows the use of the Bayesian backfitting algorithm[20] to update $(\mathcal{T}_t, \mathcal{M}_t)$ sequentially for $t = 1, \ldots, m$ by using the Metropolis-Hastings algorithm[12] to sample $\mathcal{T}_t$ using (1.3) and then sampling $\mathcal{M}_t$ from its full conditional. The Gibbs sampler,

$$\begin{aligned}
1: &\quad \mathcal{T}_1|R_1, \sigma^2 \\
2: &\quad \mathcal{M}_1|\mathcal{T}_1, R_1, \sigma^2 \\
3: &\quad \mathcal{T}_2|R_2, \sigma^2 \\
4: &\quad \mathcal{M}_2|\mathcal{T}_1, R_2, \sigma^2 \\
&\vdots \\
2m-1: &\quad \mathcal{T}_m|R_m, \sigma^2 \\
2m: &\quad \mathcal{M}_m|\mathcal{T}_m, R_m, \sigma^2 \\
2m+1: &\quad \sigma|\mathcal{T}_1, \mathcal{M}_1, \cdots, \mathcal{T}_m, \mathcal{M}_m, \mathcal{E}
\end{aligned}$$

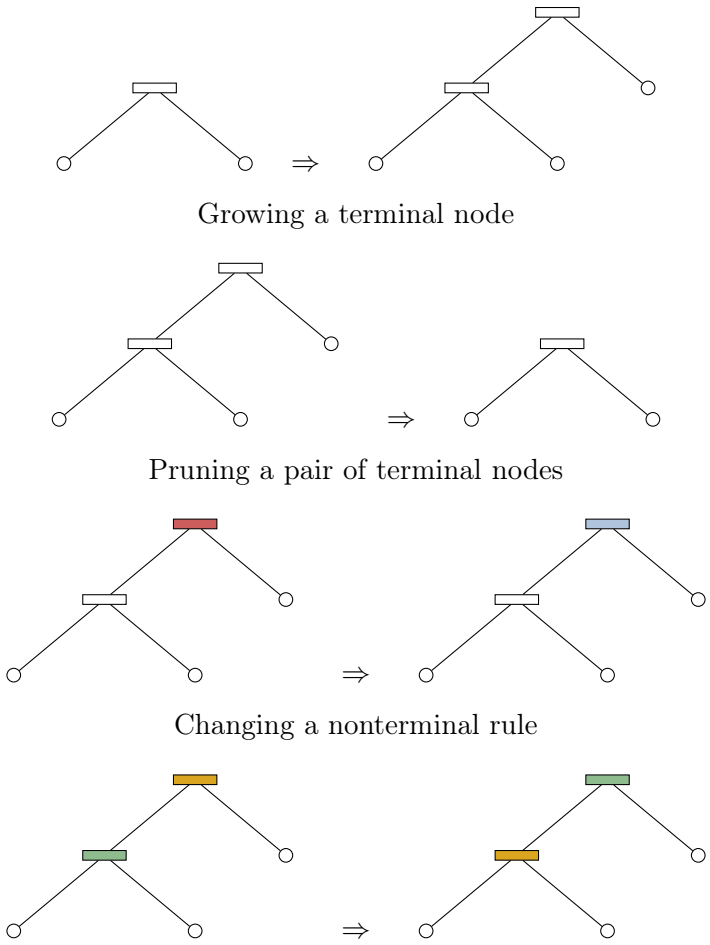

Figure 1.4: Graphical illustration of four moves of a new tree.

proceeds by proposing a change to the first tree's structure $\mathcal{T}_t$ with Metropolis-Hastings steps. Given the tree structure, samples from the posterior of the leaf parameters $mathcalM_t$ are drawn. Each tree is updated using partial residuals R_{-j}. Finally, the sample from the posterior of σ^2 is drawn form based on the full model residuals $\mathcal{E} := \boldsymbol{y} - \sum_{t=1}^{m} \mathcal{T}_t^{\mathcal{M}}(\boldsymbol{X})$ conditional on the updated set of tree structures and leaf parameters [23].

Despite the recent popularity of BART, the estimates obtained from BART have some drawbacks; in particular, BART lacks smoothness due to the distinct partitions as illustrated in Figure 1.1 (bottom-left). To overcome this lack of smoothness, Linero and Yang [29] proposed SBART, which forms predictions by averaging over many random paths down the tree. In the splitting

rule in SBART at branch b, $\boldsymbol{x}$ goes left at branch b with probability $\psi(x_{j_b}; C_b, \kappa_b) = \psi\left(\frac{x_j - C_b}{\kappa_b}\right)$, where $\kappa_b > 0$ is a bandwidth parameter controlling the sharpness of the decision. The probability associated to a particular leaf node is given by

$$\phi(x; \mathcal{T}, \ell) = \prod_{b \in A(\ell)} \psi(x_{j_b}; C_b, \kappa_b)^{1-R_b} (1 - \psi(x_{j_b}; C_b, \kappa_b))^{R_b}$$

where $A(\ell)$ is the set of ancestor nodes of leaf ℓ and $R_b = 1$ if the path for the leaf ℓ goes right at the branch b, with $R_b = 0$ otherwise. By averaging over all possible paths down the tree, each leaf from SBART has a global impact on f which allows the model to share information across different covariate regions. Figure 1.1 (bottom-right) illustrates the partition which is determined by the different splitting rule with the logistic gating function $\psi(x) = (1 + e^{-x})^{-1}$. The splitting rules of BART induce a distinct partition of the predictor space $\mathcal{X} = [0, 1]^2$. On the other hand, SBART smooths the BART partition. Such smoothness is useful when the underlying $f(\boldsymbol{x})$ is thought to be smooth. We note that if the bandwidth parameter tends 0, the splitting rule of SBART approaches the splitting rule of original BART. The parameter κ_b controls the the sharpness of the decision, with the model approaching a hard decision tree as $\kappa_b \to 0$, and approaching a constant model as $\kappa_b \to \infty$. In this dissertation, the logistic gating function $\psi(x) = (1 + e^{-x})^{-1}$ and $\tau^{(-1)} = 10$ is used for both simulation study and real data analysis.

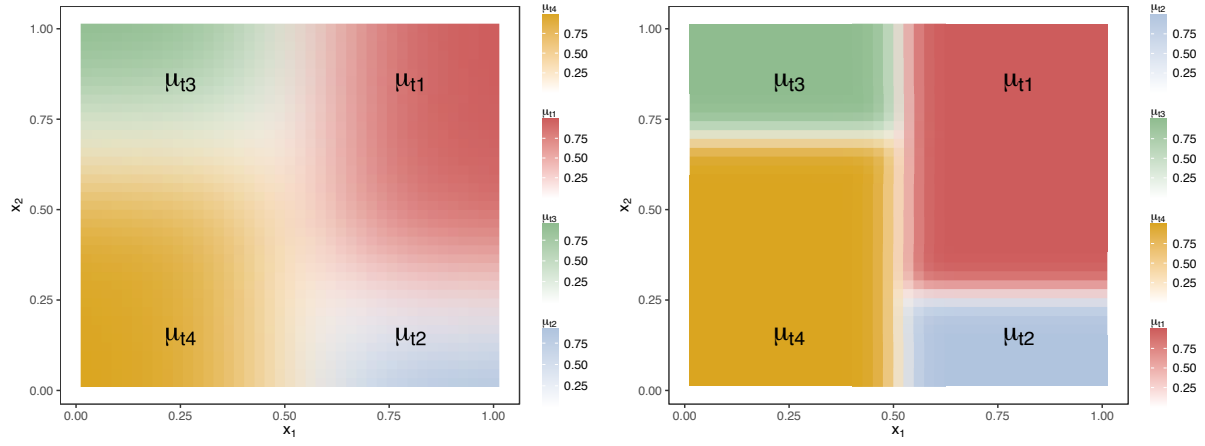

Figure 1.5: The partitions with $\tau_b^{-1} = 10$ (left) and $\tau_b^{-1} = 50$ (right) from SBART.

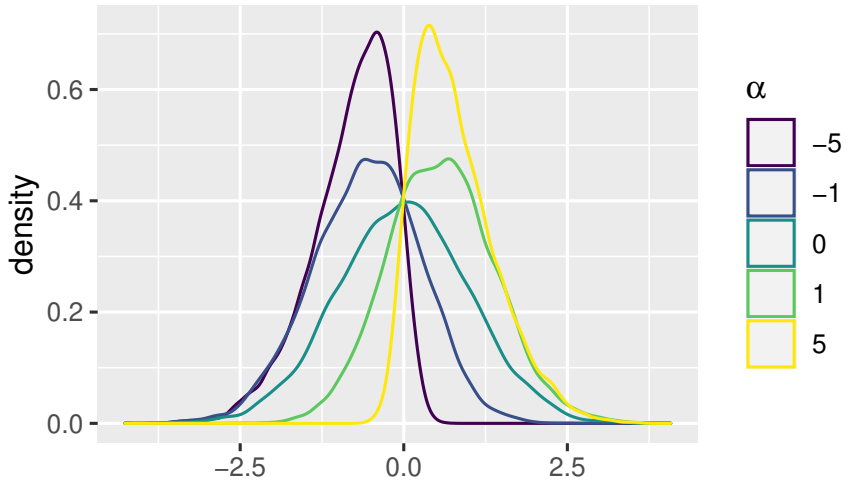

Figure 1.6: Probability density function of skew normal distribution varying skewness levels.

1.2.2 Skew-normal distribution

We write $U \sim \mathcal{SN}(\xi, \sigma^2, \alpha)$ if U follows a skew-normal (SN) distribution[6] with probability density function

$$p(u) = \frac{2}{\sigma}\phi\left(\frac{u-\xi}{\sigma}\right)\Phi\left(\alpha\frac{u-\xi}{\sigma}\right), \tag{1.4}$$

where ϕ and Φ are the standard normal density and cumulative distribution function, respectively. The parameters ξ and σ are location and scale parameters, while α is a *shape* parameter which allows for the density to exhibit skewness.

Some basic properties of the $mathcalSN(0, 1, \alpha)$ distribution given in [4] are:

1 $\mathcal{SN}(0, 1, 0) = \mathcal{N}(0, 1)$

2 If $U \sim \mathcal{SN}(0, 1, \alpha)$, then $-U \sim \mathcal{SN}(0, 1, -\alpha)$

3 As $\lambda \to \pm\infty$ tends to the half-normal distribution.

4 If $U \sim \mathcal{SN}(0, 1, \alpha)$, then $U^2 \sim \chi^2_1$

The moment generating function and the first moments of U are given in [4]

$$M_U(t) = 2\exp\left(\xi t + \frac{\sigma^2 t^2}{2}\right)\Phi(\sigma\delta t),$$
$$E(U) = \xi + \sigma\delta\sqrt{\frac{2}{\pi}}, \qquad Var(U) = \sigma^2\left(1 - \frac{2\delta^2}{\pi}\right)$$

where $\delta = \alpha/\sqrt{1+\alpha^2}$.

For ease of posterior computation, we will use the following stochastic representation of the skew-normal distribution (1.4) as

$$U \stackrel{d}{=} \lambda|Z| + W \tag{1.5}$$

where Z and W are respectively independent $\mathcal{N}(0,1)$ and $\mathcal{N}(\xi,\tau)$ random variables, with $\lambda = \frac{\alpha\sigma}{\sqrt{1+\alpha^2}}$ and $\tau = \frac{\sigma^2}{1+\alpha^2}$.

Different extensions[5, 3, 38] of the SN distribution to the multivariate setting have been proposed. For extending skew-BART to multivariate responses, we consider the generalized multivariate skew-normal (MSN) distribution [1], which is both practical and flexible. Let $\mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^k$ and $k \times k$ positive definite covariance $\boldsymbol{\Sigma}$. A $k \times 1$ random vector $\boldsymbol{y}$ follows an MSN distribution with a $k \times k$ skewness matrix $\boldsymbol{\Delta} = \text{diag}(\boldsymbol{\lambda})$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T \in \mathbb{R}^k$, denoted by $\boldsymbol{y} \sim \mathcal{SN}_k(\boldsymbol{\mu}, \Sigma, \Delta)$, if its probability density function is given by

$$2^k \phi_k(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Omega})\Phi_k(\boldsymbol{\Delta}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}); \boldsymbol{0}, \boldsymbol{\Lambda}^{-1}), \tag{1.6}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Delta}\boldsymbol{\Delta}^T$ and $\boldsymbol{\Lambda} = \boldsymbol{I}_k + \boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}$. Note that (1.6) belongs to the class of fundamental skew distributions considered by Arellano-valle and Genton [3]. From Proposition 1 of Arellano-Valle et al. [1], $\boldsymbol{y}$ can be represented stochastically as

$$\boldsymbol{y} \stackrel{d}{=} \boldsymbol{\Delta}|\boldsymbol{Z}| + \boldsymbol{U} \ , \tag{1.7}$$

where $\boldsymbol{Z} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{I}_k)$ and $\boldsymbol{U} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are independent k-variate normal random vectors. Here, $|\boldsymbol{Z}|$ denotes the element-wise absolute value of $\boldsymbol{Z}$. The family of MSN densities has the desirable and practically useful property of being closed under marginalization. Additionally, each of the individual components of the MSN distribution has a univariate skew-normal marginal. Sahu et al.

[38] and Bhingare et al. [9] present physical interpretations of univariate SN density of (1.5), and multivariate SN density of (1.7), in terms of the independent skewing shocks $\boldsymbol{Z}$.

A direct consequence of 1.7, related with the moments of the skew-normal random vector, is given

$$E[\boldsymbol{Y}] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\delta} \quad \text{and} \quad V[\boldsymbol{Y}] = \boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\Delta}^2.$$

1.2.3 The negative-binomial model

Let y_i denote an count response for individual i where $i = 1, \cdots, N$. The common choice for modeling count data is Poisson model, $y_i \sim Pois(\lambda)$. However, Poisson model is limited by the fact that mean and variance are both equal to λ. This assumption does not hold in many applications, where responses are often over-dispersed. Also, the Poisson models is difficult to perform fully Bayesian inference under Gaussian prior since it has no tractable representation. To relax this assumption, we consider the negative binomial(NB) distribution $y_i \sim NB(r, p)$ generated by a hierarchical Poisson model. Suppose

$$(y_i|\lambda_t) \sim Pois(\lambda_i)$$
$$(y_i|\xi, \psi_i) \sim Gamma(\xi, e^{\psi_i})$$

By marginalizing for λ_i, a negative-binomial distribution for y_i is generated:

$$p(y_i) = \int_0^\infty \text{Pois}(y_t; \lambda_i) Gamma\left(\lambda_i; r, ^{\psi_i}\right) d\lambda = \frac{\Gamma(r + y_i)}{y_i!\Gamma(r)}(1 - p_i)^{\xi} p_i^{y_i},$$

where $\Gamma(\cdot)$ is the gamma function, $r \geq 0$ is dispersion parameter and $p_i = e^{\psi_i}/(1 + e^{\psi_i})$ is a probability parameter. The NB distribution is also known as the gamma-Poisson distribution because of gamma distribution prior on Poisson distribution. Since the variance $\xi e^{\psi_i}(1 + e^{\psi_i})$ is larger than the mean ξe^{ψ_i} in negative binomial distribution, it is more favorable than the Poisson distribution for modeling overdispersed count data.

The Negative binomial regression is commonly performed with the maximum likelihood estimator (MLE). Although Bayesian approaches are favored to model the uncertainty of estimation and to incorporate prior information, the limitation of developing simple and efficient algorithms for posterior computation has made Bayesian analysis of counts underdeveloped. However, recent latent

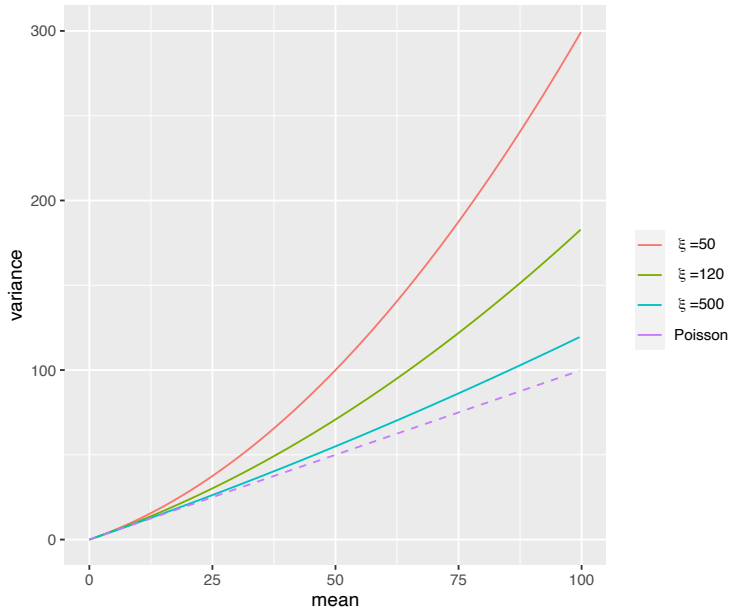

Figure 1.7: Relationship between mean and variance for different settings of shape parameter ξ, illustrating variability of the NB model and Poisson model.

variable construction based on Polya-Gamma random variables makes Bayesian inference appealing in the negative-binomial model with efficient sampling derived by exploiting conditional conjugacy [35]. One of the biggest benefit on the construction is that the negative binomial likelihood can be represented as a mixture of normals with Polya-Gamma mixing distribution.

A random variable X has a Polya-Gamma distribution with parameters $b > 0$ and $c \in \mathbb{R}$, denoted $X \sim PG(b, c)$, if

$$X \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}$$

where the $g_k \sim Gamma(b, 1)$ are independent gamma random variable and where $\overset{D}{=}$ indicates equality in distribution. Although the density of a Polya-Gamma random variable includes an infinite series, its expected value is derived with closed form:

$$E(\omega) = \frac{b}{2c} \tanh(c/2).$$

Denote $p(\omega)$ the density of the random variable $\omega \sim \mathrm{PG}(b, 0)$ for $b > 0$, then for all $a \in \mathbb{R}$

$$\frac{\left(e^{\psi}\right)^{a}}{\left(1+e^{\psi}\right)^{b}} = 2^{-b} e^{\kappa\psi} \int_{0}^{\infty} e^{-\omega\psi^{2}/2} p(\omega) d\omega \tag{1.8}$$

where $\kappa = a - b/2$. The negative binomial likelihood can be rewritten with the integral identity for $b > 0$,

$$(1-p_t)^{\xi} p_t^{y_t} = \frac{\{\exp(\psi_t)\}^{y_t}}{\{1+\exp(\psi_t)\}^{h+y_t}} \propto e^{\kappa_t \psi_t} \int_{0}^{\infty} e^{-\omega_t \psi^{2}/2} p(\omega \mid \xi + y_t, 0)\, d\omega$$

where $\kappa_t = (y_t + \xi)/2$, and where the mixing distribution is Polya-Gamma. The quadratic form Q allows to be conditionally conjugate to any Gaussian or mixture-of-Gaussians prior for ψ_t conditional upon ω_t. This exploiting conditional conjugacy enables to develop efficient sampling.

Also, the conditional distribution

$$p(\omega \mid \psi) = \frac{e^{-\omega\psi^{2}/2} p(\omega)}{\int_{0}^{\infty} e^{-\omega\psi^{2}/2} p(\omega) d\omega}$$

is also in the Polya-Gamma class: $(\omega \mid \psi) \sim \mathrm{PG}(b, \psi)$. In this sense, the Polya-Gamma distribution is conditionally conjugate to the NB likelihood, which is very useful for Gibbs sampling.

Despite the complicated form of the density function, it is relatively easy to simulate random Polya-Gamma draws with the above described facts allowing straightforward Bayesian inference for negative-binomial models.

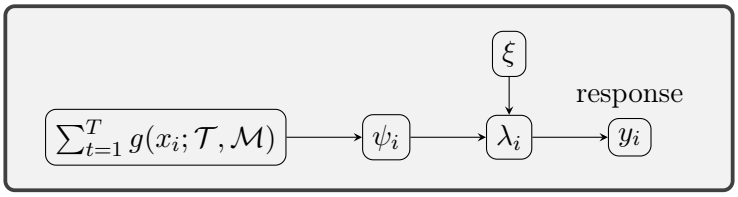

Figure 1.8: Representations of the negative-binomial (NB) regression model. Graphical model for standard gamma-Poisson mixture representation of the NB. The sum of trees defines the scale parameter for a gamma r.v. with shape parameter ξ, giving $\lambda_t \sim \mathrm{Gam}\left(e^{\psi_t}, \xi\right)$ which is in turn the rate for a Poisson spike count: $y_i \sim \mathrm{Pois}(\lambda_i)$.

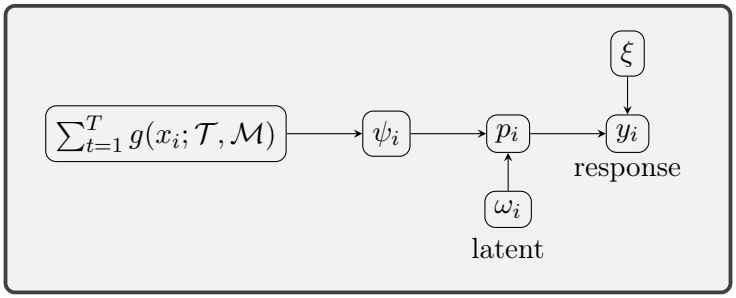

Figure 1.9: Graphical model illustrating novel representation as a Polya-Gamma (PG) mixture of normals. The counts are represented as NB distributed with shape ξ and rate $p_i = 1/\left(1 + e^{-\psi_i}\right)$. The latent variable ω_i is conditionally PG, while ψ are normal given (ω_i, ξ), which facilitates efficient inference.

CHAPTER 2

BAYESIAN ADDITIVE REGRESSION TREES FOR MULTIVARIATE SKEWED RESPONSES

2.1 The `skewBART` and `multi-skewBART` models

Highly skewed multivariate responses and commonly observed in many biomedical and clinical research problems. For example, in a preliminary analysis (see Figure 2.1) of periodontal disease (PD) status [15] among a population of Type-2 diabetic Gullah-speaking African-Americans (henceforth, the GAAD study) two popular PD endpoints — the mean periodontal pocket depth (PPD) and mean clinical attachment level (CAL) — are highly skewed both marginally and jointly. The data analytic goal here is to assess the cross-sectional associations of important covariates, say, age and diabetes level, on the bivariate responses, i.e., with the average/mean PPD and average/mean CAL of all measured tooth-sites of the participants.

To model the relationship between a multivariate response and a collection of covariates without imposing any pre-determined functional form, we consider the nonparametric regression of scalar responses $Y_1, \ldots, Y_n$ on the corresponding p-dimensional covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, given by

$$Y_i = f(\boldsymbol{x}_i) + \epsilon_i \ , \tag{2.1}$$

where the errors $\epsilon_1, \ldots, \epsilon_n$ are iid with common density $h_\epsilon(\cdot)$. A popular choice for modeling the unknown function $f : \mathbb{R}^p \to \mathbb{R}$ is to use an ensemble of decision trees, such as random forests [10], or boosted decision trees [16]. In this dissertation, we focus on the Bayesian additive regression trees (BART) model introduced by Chipman et al.[12]. Unlike the parametric regression functions utilized to model skewed multivariate responses [9], decision tree ensembles efficiently capture both nonlinear and interaction effects of covariates in $f(\boldsymbol{x})$, automatically. Additionally, compared to other machine learning algorithms, BART models provide a number of appealing advantages. First, they are typically robust to the choice of tuning parameters, and yield high prediction accuracy. Second, unlike competing algorithms often lacking well-defined uncertainty quantification around their predictions, BART approximates the full posterior equipped with a highly effective choice

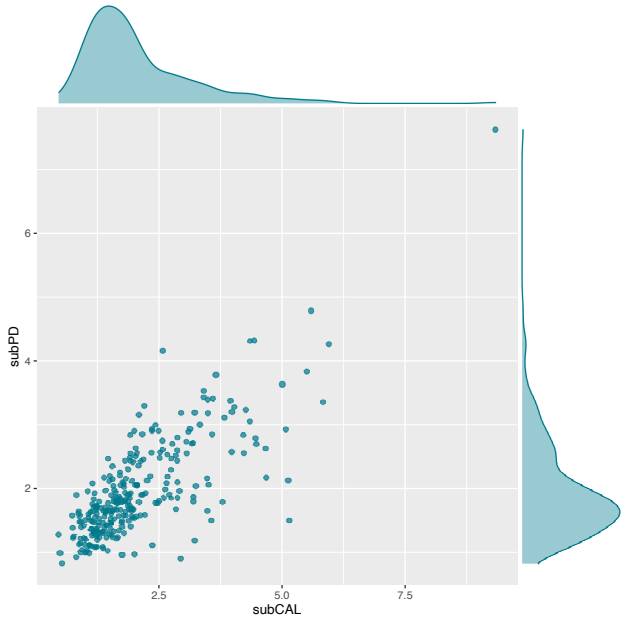

Figure 2.1: GAAD data: Scatter plot of PPD and CAL responses, averaged for each subject, with marginal density estimates.

of priors and hyperparameters. Recent work has also established attractive theoretical properties for BART [37, 29]. BART models have successfully been applied to many statistical problems, including classification [31], survival analysis [39, 8], density estimation [26], and high dimensional sparse regression [27].

However, the available popular versions of BART models come with various limitations. For example, in studies with highly skewed responses, BART with the error density $h_\epsilon(\cdot)$ being a Gaussian density may not be appropriate. Bhingare at al.,[9] reveals that the mean PPD and mean CAL are highly skewed, and methods which adapt to skewed responses tend to outperform methods based on Gaussian errors regarding precision of the estimated parameters and predictions of future outcomes. Second, as illustrated by Linero and Yang [29], the estimate of $f(\cdot)$ obtained from the usual BART models lack smoothness, leading to suboptimal performance when $f(\boldsymbol{x})$ varies smoothly with $\boldsymbol{x}$. To improve performance of the estimate when true $f(\cdot)$ is believed to be smooth, Linero and Yang[29] introduced the soft BART (SBART) model.

In this dissertation, we introduce the `skewBART` model, which takes the error density $h_\epsilon(\cdot)$ in equation (2.1) to be a skewed normal density to accommodate non-Gaussian responses. This model

also extends the SBART model to obtain smooth estimates of $f(\cdot)$ under skewed errors. We develop an efficient Gibbs sampler based on the Bayesian backfitting algorithm[12] to perform practical nonparametric Bayesian inference.

In addition to accommodating skewness of the error, we introduce a skewed multivariate response model, with multivariate decision trees for regression function and a multivariate skewed distribution for the error vector for each subject. The use of multivariate decision trees has been used both to handle mixed-type responses such as zero-inflated log-normal/gamma distributions[28], as well as to implement *targeted smoothing* over space or time[40]. To the best of our knowledge, this is the first attempt to bring multivariate decision tree methods to bear in the BART framework to model multivariate skewed response data. To increase the precison of the inference, our `multi-skewBART` takes into account the association of the multivariate responses within each subject, borrowing information across the different responses by using the same decision trees for the mean vector. We further illustrate the advantages of our `skewBART` and `multi-skewBART` proposals in practice on the GAAD study.

The rest of the paper is organized as follows. After a brief review of both the BART and SBART models as well as the important properties of the skew normal distribution, Section **??** introduces the `skewBART` and `multi-skewBART` models, respectively for univariate and multivariate responses. Details on the Bayesian inference, including prior specifications and the related Markov chain Monte Carlo(MCMC) algorithm for posterior computation are outlined in Section 3.2. In Section 2.3, we compare the finite sample performances (including out-of-sample predictions) of our proposals to existing alternatives using synthetic data. In Section 2.4, we demonstrate the application to the motivating GAAD oral health dataset. Section 2.5 concludes with a discussion. Details on the proposed MCMC algorithm appear in the Appendix.

2.1.1 `skewBART`: univariate

We first introduce the univariate `skewBART` model, which is an extension of BART to accommodate skewed responses. We approximate the nonparametric regression function $f(\cdot)$ in (2.1) using a sum-of-trees model with an SN distribution for $Y = (Y_1, \ldots, Y_n)$ as

$$Y_i = \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{SN}(0, \sigma^2, \alpha). \tag{2.2}$$

Using the stochastic representation of (1.5), we can rewrite (2.2) as

$$Y_i = \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \lambda|Z_i| + W_i \tag{2.3}$$

where $W_i \sim \mathcal{N}(0, \tau)$, $Z_i \sim \mathcal{N}(0, 1)$, $\lambda = \alpha\sigma/\sqrt{1+\alpha^2}$ and $\sqrt{\tau} = \sigma/\sqrt{1+\alpha^2}$. In Section 2.2.1, the reparameterization (2.3) allows us to construct a simple data augmentation step to fit the `skewBART` model (2.2). For the univariate model, we use the SBART framework of Linero and Yang [29] to adapt to smoothness in $f(\cdot)$.

2.1.2 `skewBART`: multivariate

We extend the sum-of-trees model (2.2) using the MSN distribution (1.7) to accommodate multivariate outcomes. This extension improves prediction accuracy by incorporating the correlation between responses, as well as the dependency in error estimation. Our proposed multivariate `skewBART`, called `multi-skewBART`, approximates the multivariate response $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^T \in \mathbb{R}^{n\times k}$ as

$$\boldsymbol{Y}_i = \sum_{t=1}^{m} \boldsymbol{g}_t^*(\boldsymbol{x}_i; \mathcal{T}_t, \boldsymbol{M}_t) + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \overset{iid}{\sim} \mathcal{SN}_k(\boldsymbol{0}, \Sigma, \Delta), \tag{2.4}$$

where $\boldsymbol{g}_t^*(\boldsymbol{x}_i; \mathcal{T}_t, \boldsymbol{M}_t)$ returns a k-dimensional $\boldsymbol{\mu}_{t\ell}$ if $\boldsymbol{x}_i$ is associated with the leaf node ℓ in $\mathcal{T}_t$ for $\ell = 1, \ldots, n_t$. The tree $\mathcal{T}_t$ in (2.4) has a binary tree structure like univariate `skewBART` but $\boldsymbol{M}_t = \{\boldsymbol{\mu}_{t1}, \cdots, \boldsymbol{\mu}_{tn_t}\}$ is now a set of $k-$dimensional leaf parameters. Using the stochastic representation of the MSN distribution in (1.7), the `multi-skewBART` model of (2.4) is now expressed as

$$\boldsymbol{Y}_i = \sum_{t=1}^{m} \boldsymbol{g}_t^*(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \Delta|\boldsymbol{Z}_i| + \boldsymbol{W}_i \ ,$$

where $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{ik}) \overset{iid}{\sim} \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{I}_k)$, $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{ik}) \overset{iid}{\sim} \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, Σ is a positive definite $(k \times k)$ scale matrix, and the skewness matrix Δ is a diagonal $(k \times k)$ matrix with diagonal entries $\lambda_1, \ldots, \lambda_k$.

2.2 Prior specification and posterior computation

In the literature it has been seen empirically that the default priors proposed by Chipman et al. [12] work remarkably well in practice. For the tree structures, we use the default prior for the BART model and provide priors for rest of the parameters in h_ϵ. We then derive Gibbs samplers for fitting the models.

2.2.1 `skewBART`: prior choices and MCMC computation

Each column of the design matrix X is distributed approximately uniformly on $[0,1]$ by independent quantile normalizations. Additionally, following Chipman et al.[12], the dependent variable Y is standardized by first shifting and rescaling. Then, the leaf parameters $\mu_{t\ell}$ are assumed independent and identically distributed as $\mathcal{N}(0, \sigma_\mu^2/m)$. With this prior, the leaf parameters $\mu_{t\ell}$ are shrunk towards zero, thus regularizing the effect of the individual tree components to contribute only a small part in the overall fit.

For choosing the bandwidth κ_b, Linero and Yang [29] recommend using tree specific κ_t's shared across branches in a fixed tree, with $\kappa_t \sim \text{Exp}(0.1)$. We specify a half-Cauchy prior $\sqrt{\tau} \sim \text{Cauchy}_+(0, \hat{\tau})$, where $\hat{\tau}$, where $\hat{\tau}$ is chosen empirically. We obtain $\hat{\tau}$ by fitting the lasso using the `glmnet` package in `R`. We use a conjugate univariate normal $\mathcal{N}(0, \delta)$ prior for λ to allow both positive and negative skewness in (2.2), where, a large value of δ corresponds to a non-informative prior opinion about the amount of skewness. Note that the full conditional distribution of Z depends on both λ and τ. In summary, our proposed prior is given by

$$
\begin{aligned}
&Y_i = \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \lambda |Z_i| + W_i, \\
&\lambda \sim \mathcal{N}(0, \delta), && Z_i \sim \mathcal{N}(0,1), \\
&W_i \sim \mathcal{N}(0, \tau), && \sqrt{\tau} \sim \text{Cauchy}_+(0, \hat{\tau}) \\
&\alpha = 0.95, && \beta = 2.
\end{aligned}
$$

Below, we describe our MCMC algorithm for fitting `skewBART` via data augmentation. Our strategy is to augment the latent variables $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ within the Bayesian backfitting algorithm[20] in order to sample approximately from the posterior distribution

$$\pi((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \lambda, \tau, \boldsymbol{Z} \mid \boldsymbol{Y})$$

using MCMCs, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. By abuse of notation, we use the symbol $[\theta \mid \gamma]$ to denote the conditional posterior distribution of θ given γ. Our Gibbs sampler draws $(\mathcal{T}_j, \mathcal{M}_j)$ for $j = 1, \ldots, m$ given $(\mathcal{T}_{-t}, \mathcal{M}_{-t}, \boldsymbol{Z}, \lambda, \tau)$ from

$$[\mathcal{T}_j, \mathcal{M}_j | \mathcal{T}_{-j}, \mathcal{M}_{-j}, \boldsymbol{Z}, \boldsymbol{Y}, \lambda, \tau], \tag{2.5}$$

where $\mathcal{M}_{-j} = \{\mathcal{M}_t : t \neq j\}$ and $\mathcal{T}_{-j} = \{\mathcal{T}_t : t \neq j\}$. Then, samples of τ, λ and $\boldsymbol{Z}$ are drawn from conditional posterior distributions

$$
\begin{aligned}
&[\tau|\mathcal{T}_1, \ldots, \mathcal{T}_m, \mathcal{M}_1, \ldots, \mathcal{M}_m, \boldsymbol{Y}, \boldsymbol{Z}, \lambda], \\
&[\lambda|\mathcal{T}_1, \ldots, \mathcal{T}_m, \mathcal{M}_1, \ldots, \mathcal{M}_m, \boldsymbol{Y}, \boldsymbol{Z}, \tau], \\
&[\boldsymbol{Z}|\mathcal{T}_1, \ldots, \mathcal{T}_m, \mathcal{M}_1, \ldots, \mathcal{M}_m, \boldsymbol{Y}, \lambda, \tau].
\end{aligned} \tag{2.6}
$$

To implement the m draws of $(\mathcal{T}_j, \mathcal{M}_j)$ in (3.3), the draws are equivalent to sampling from

$$
[\mathcal{T}_j, \mathcal{M}_j|\boldsymbol{R}_j, \lambda, \tau] \ , \tag{2.7}
$$

where $\boldsymbol{R}_j = \boldsymbol{Y} - \sum_{k \neq j} g(X; \mathcal{T}_k, \mathcal{M}_k) - \lambda|\boldsymbol{Z}|$. This allows us to use the existing Bayesian backfitting algorithm to update the $(\mathcal{T}_j, \mathcal{M}_j)$ for $j = 1, \ldots, m$. The conditional posterior of λ given rest of the parameters, $[\lambda|\mathcal{T}_1, \ldots, \mathcal{T}_m, \mathcal{M}_1, \ldots, \mathcal{M}_m, \boldsymbol{Y}, \boldsymbol{Z}, \tau]$, has a form similar to the posterior of the regression parameters for the linear regression model of $\boldsymbol{Y}$ on $|\boldsymbol{Z}|$ under a Gaussian prior for λ; hence, this conditional posterior is also Gaussian. The full conditionals of the components of $|\boldsymbol{Z}|$ can be derived in closed form as a collection of independent truncated Gaussian random distributions. The Bayesian backfitting algorithm for `skewBART` is described in the Algorithm 3 below, with the exact conditional posterior distributions corresponding to (3.4) relegated to the Appendix.

2.2.2 `multi-skewBART`: prior choices and MCMC computation

To extend the univariate `skewBART` model to handle multivariate k-dimensional responses, we need to replace the univariate priors with priors on vector/matrix parameters. Define $\tilde{\boldsymbol{Y}} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ and $\tilde{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ik})$ and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ik})$. For $\pi(\boldsymbol{\mu}_{t\ell} \mid \mathcal{T}_t)$, we use the conjugate multivariate normal distribution, $\mathcal{N}_k(\boldsymbol{\mu}_M, \Sigma_M)$, that allows for (1.3) to be computed in closed form as in the univariate BART model. The hyper parameters $\boldsymbol{\mu}_M$ and Σ_M are chosen to assign high probability to the range $(\boldsymbol{Y}_{\min}, \boldsymbol{Y}_{\max})$, where the j'th row of this matrix is given by $(\min_i Y_{ij}, \max_i Y_{ij})$. We use a data-informed estimate of Σ to center the prior distribution as we did in the univariate case. We assign an inverse-Wishart prior distribution to Σ. To update $\boldsymbol{\Delta} = \text{diag}(\boldsymbol{\lambda})$, we use a conjugate multivariate normal prior $\mathcal{N}_k(\boldsymbol{0}, \Lambda)$. We summarize our prior choices below:

$$
\boldsymbol{Y}_i = \sum_{t=1}^{m} \boldsymbol{g}_t^*(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \Delta|\boldsymbol{Z}_i| + \boldsymbol{W}_i
$$

Algorithm 1 A single iteration of the Bayesian backfitting algorithm (`skewBART`)

for $t = 1$ to m **do**
  Set $\boldsymbol{R}_t \leftarrow \boldsymbol{Y} - \sum_{k \neq t} g(X; \mathcal{T}_k, \mathcal{M}_k) - \lambda \boldsymbol{Z}$
  Propose a new tree $\mathcal{T}_t^* \sim Q(\mathcal{T}_t \to \mathcal{T}_t^*)$
  Compute the acceptance ratio

$$r = \frac{\Lambda(\mathcal{T}_t^*) Q(\mathcal{T}_t^* \to \mathcal{T}_t)}{\Lambda(\mathcal{T}_t) Q(\mathcal{T} \to \mathcal{T}_t^*)}$$

  where $\Lambda(\mathcal{T}_t) = L(\mathcal{T}_t \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}) \times \pi(\mathcal{T}_t)$.
  Accept $\mathcal{T}_t^*$ if $U \leq r$ where $U \sim \text{Uniform}(0, 1)$, otherwise reject it.
  Sample κ_t using Metropolis-Hastings, sampling the proposal from the prior distribution.
  Update the leaf node parameters.
end for
for $i = 1$ to N **do**
  Sample Z_i from the full conditional distribution of Z_i.
end for
Sample λ from the full conditional distribution of λ.
Sample τ from the full conditional distribution of τ.

$$\boldsymbol{\lambda} \sim \mathcal{N}_k(0, \Lambda), \qquad \boldsymbol{Z}_i \sim \mathcal{N}(0, \boldsymbol{I}_k),$$
$$\boldsymbol{W}_i \sim \mathcal{N}_k(0, \Sigma), \qquad \Sigma \sim \text{inverse-Wishart}(\nu_0, S_0^{-1}),$$
$$\alpha = 0.95, \qquad \beta = 2.$$

The Gibbs sampler draws $(\mathcal{T}_j, \boldsymbol{M}_j)$ given $(\mathcal{T}_{-t}, \boldsymbol{M}_{-t}, \boldsymbol{Z}, \Delta, \Sigma)$ for $j = 1, \ldots, m$ from

$$[\mathcal{T}_j, \boldsymbol{M}_j | \tilde{\boldsymbol{R}}_j, \Delta, \Sigma] \ ,$$

where $\tilde{\boldsymbol{R}}_j = \tilde{\boldsymbol{Y}} - \sum_{k \neq j} \boldsymbol{g}^*(X; \mathcal{T}_k, \boldsymbol{M}_k) - \Delta |\tilde{\boldsymbol{Z}}|$. Then, we sample of Δ, Σ and $\boldsymbol{Z}$ from each full conditional distribution given as

$$[\Delta | \mathcal{T}_1, \ldots, \mathcal{T}_m, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_m, \boldsymbol{Y}, \boldsymbol{Z}, \Sigma],$$
$$[\Sigma | \mathcal{T}_1, \ldots, \mathcal{T}_m, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_m, \boldsymbol{Y}, \boldsymbol{Z}, \Delta],$$
$$[\boldsymbol{Z} | \mathcal{T}_1, \ldots, \mathcal{T}_m, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_m, \boldsymbol{Y}, \Delta, \Sigma].$$

The MCMC algorithm for this model is similar to the original MCMC scheme for the univariate skew-BART, but now extended to the multivariate setting.

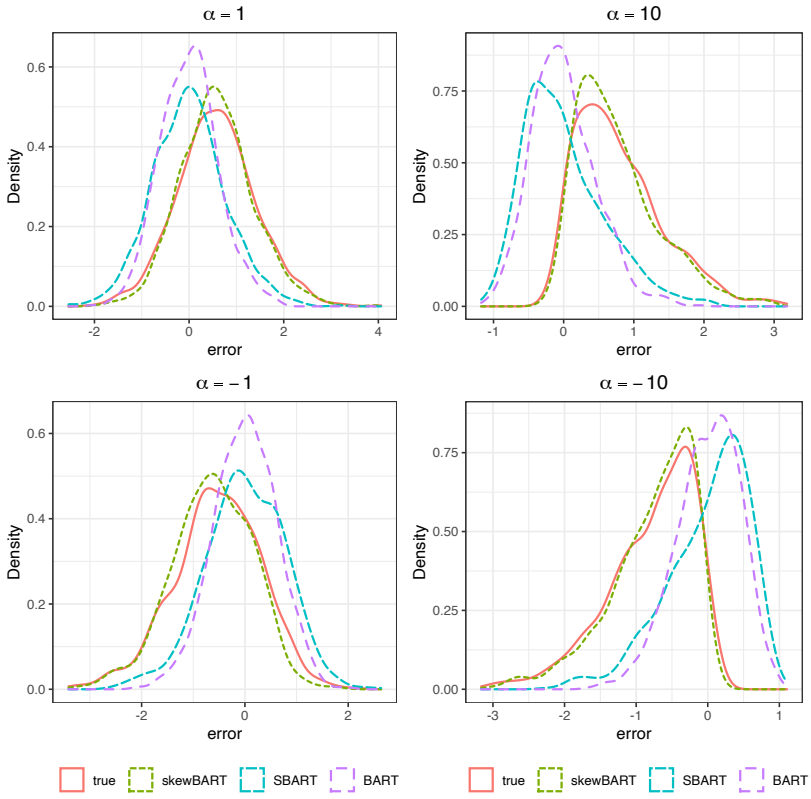

Figure 2.2: Simulation study (univariate): Error density histograms corresponding to four different skewness parameters, $\alpha = (-1,\ 1,\ -10,\ 10)$, overlaid with the density histograms from the `skewBART`, SBART and BART fits.

2.3 Simulation study

In this section, we conduct simulation studies to compare `skewBART` and `multi-skewBART` to existing methods. The simulations are aimed to evaluate the estimated regression function along with out-of-sample predictive performances.

2.3.1 Univariate responses

We first examine the inferences obtained by `skewBART` compared to BART and SBART, when the true error distribution takes on four different skewness levels. For each replicated data-set, we simulate $n = 250$ observations using the nonparametric regression model of (2.1) with the known population mean function

$$\boldsymbol{Y}_i \ = \ 10\sin(\pi X_{i1} X_{i2}) + 20(X_{i3} - 0.5)^2 + 10X_{i4} + 5X_{i5} + \epsilon_i \tag{2.8}$$

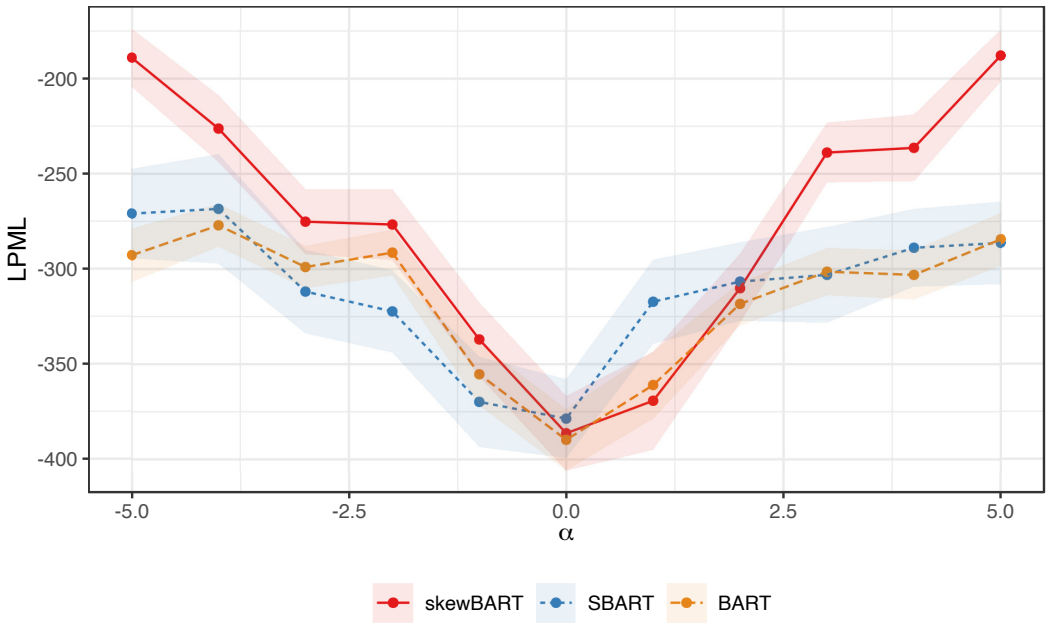

Figure 2.3: Simulation study (univariate): LPML, with shaded bars representing ±2 standard error, for varying $\alpha \in (-0.5, 0.5)$, with increments of 1.

introduced by Friedman [17], where $X_i \sim \text{Uniform}([0,1]^5)$ and $\epsilon_i \overset{iid}{\sim} \mathcal{SN}(0,1,\alpha)$. For each experiment, we use 200 trees, 5000 MCMC samples, and a burn-in of 2500 draws. Figure 2.2 presents density estimates of the distribution of the residuals from fitting the three competing methods overlaid with the true error density function, corresponding to the various choices of α, the skewness parameter. While the two plots on the left (upper and lower) panels corresponds to moderate skewness, i.e., $\alpha = -1$ and $\alpha = +1$, the two on the right (upper and lower) panels represent high magnitude of skewness, i.e., $\alpha = -10$ and $\alpha = +10$. Under both moderate and heavy skewness, the empirical distribution of the residuals from our proposed `skewBART` model appears closest to the true error distribution, revealing superior performance of the `skewBART`.

Next, we compare `skewBART` to BART as well as SBART when data generation process varies with skewness levels. We simulate $n = 250$ observations, with $\sigma^2 = 1$ and use $m = 200$ trees. The skewness values are equally spaced points in the range $(-5, 5)$, with increment of 1. We set hyperparameters for the priors of trees in BART and SBART, following recommendations by Chipman et al[12], and Linero and Yang[29]. To compare model performance we use Conditional Predictive Ordinate (CPO)[18], where $\text{CPO}_i = f(Y_i|\boldsymbol{Y}_{-i}, \boldsymbol{X})$ is the predictive density of the ith observation

given $\boldsymbol{Y}_{-1} = (Y_i, \ldots, Y_{i-1}, Y_i, \ldots, Y_n)$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. A natural summary statistic of the CPO_i's is the log pseudomarginal likelihood (LPML), given by $\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i)$. The computations of LPML are conveniently based on the Markov chain output using the `loo` package in `R`. Results corresponding to each skewness level are presented in Figure 2.3. Larger values of LPML indicate a better fit of the model.

Among the methods considered, `skewBART` performs the best, demonstrating improvement over SBART as skewness increases. For sufficiently large $|\alpha|$ we observe a substantial increase in the LPML for the `skewBART`, since the `skewBART` can capture the excess skewness through the non-Gaussian error assumption. Also, the performances of the `skewBART` and SBART are hardly distinguishable under low to negligible skewness, implying the `skewBART` does not suffer in terms of performance when the responses are not skewed.

2.3.2 Bivariate responses

Table 2.1: Simulation study (bivariate): LPML corresponding to the fits of the `multi-skewBART` and Multi-BART, for various settings of λ_1, λ_2 and ρ.

		`multi-skewBART`	Multi-BART
$\boldsymbol{\lambda} = (0, 3)$	$\rho = 0$	-356.677	-360.145
	$\rho = 0.5$	-330.328	-346.016
	$\rho = 0.9$	-287.791	-308.521
$\boldsymbol{\lambda} = (2, 3)$	$\rho = 0$	-396.685	-427.258
	$\rho = 0.5$	-382.333	-412.068
	$\rho = 0.9$	-386.478	-400.468
$\boldsymbol{\lambda} = (-2, 2)$	$\rho = 0$	-369.528	-396.559
	$\rho = 0.5$	-312.813	-384.575
	$\rho = 0.9$	-311.486	-325.217

In this section, we examine the benefits of adapting skewness for bivariate responses using synthetic data. For $\boldsymbol{Y} \in \mathbb{R}^2$, we generate data from the model:

$$\boldsymbol{Y}_i = \boldsymbol{f}^*(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i,$$

where, $\boldsymbol{x}_i \sim \text{Uniform}([0,1]^5)$ and $\boldsymbol{\epsilon}_i \sim \mathcal{SN}_2(\boldsymbol{0}, \Sigma, \Delta)$, with the scale matrix Σ and skewness matrix Δ. The function $\boldsymbol{f}^*$ returns 2-dimensional vectors corresponding to $\boldsymbol{x}_i$. We also use the Friedman's

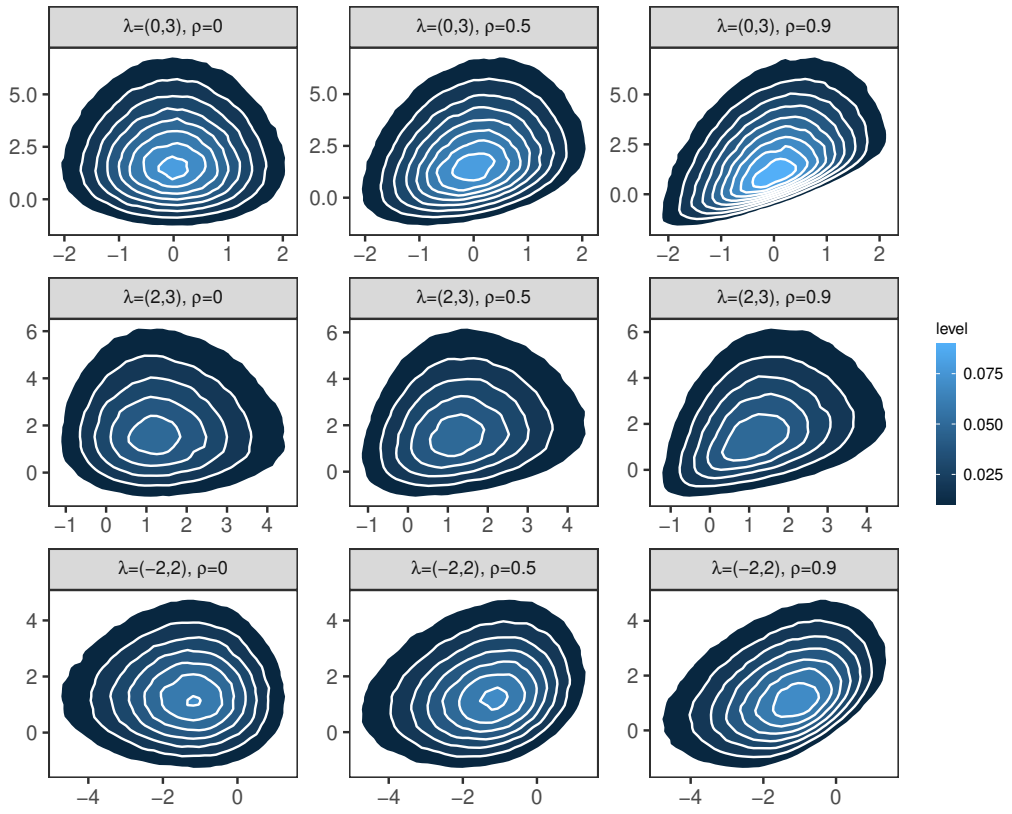

Figure 2.4: Simulation study (bivariate): The contour plots of the bivariate skew-normal distribution for $\boldsymbol{\mu} = (0, 0)$ and $\sigma_1 = \sigma_2 = 1$ with different values of (λ_1, λ_2) and ρ.

example for $\boldsymbol{f}^*$ with $N = 250$ and $\sigma_1 = \sigma_2 = 1$. Following Arellano-valle et al[2], we consider 9 different settings varying with λ_1, λ_2, the skewness parameters, and ρ, the correlation parameter in the bivariate specification, as displayed in Figure 2.4. We compare `multi-skewBART` to the multi-BART, the multivariate version of the standard BART model, via LPML, with 200 trees and 5,000 MCMC draws. Results of this simulation are presented in Table 2.1. The result shows that `multi-skewBART` outperforms multi-BART with larger LPML under all settings. It appears that both skewness and correlation have an impact on the performance. Overall, higher correlation leads to higher LPML, with the highest LPML attained under $\boldsymbol{\lambda} = (0, 3)$, with fixed correlation for both models. Thus, we conclude that the `multi-skewBART` fit is better, and its predictive performance remains highly competitive, regardless of the magnitude of skewness and correlation.

2.4 Applications: GAAD study

In the motivating GAAD study, the PD status of each subject is assessed via two correlated biomarkers, the PPD and CAL, measured at six pre-specified tooth sites of each available tooth, excluding the third molars. For our analysis, we consider the bivariate response for each subject to be $(\text{PPD}_i, \text{CAL}_i)$, where we use the *average* of the PPD and CAL across all available sites and all available teeth as the outcomes. Also available in the dataset are various subject-level covariates, such as age (in years), body mass index (BMI, in kg/m^2), glycated haemoglobin level (HbA1c, in %), gender (1 for female, 0 for male), and smoking status (1 for past or present smoker, 0 for never). We analyze 288 subjects with complete covariate information. Our data shows that female subjects are predominant (about 76%), which is not uncommon in the Gullah population[22]. The mean age is about 55 years (ranging from 26 to 87 years), with smokers constituting 31% of all subjects. About 68% of subjects are obsese (defined as BMI ≥ 30), and 60% of subjects have poorly controlled Type-2 diabeters (Hba1c ≥ 7). In contrast to previous approaches that considered parametric regression functions[36] and semiparametric formulations with skewed errors[9], our approach incorporates BART-based nonparametric regression to model the unknown regression function and to capture both nonlinear and interaction effects of the covariates.

First, we present a univariate analysis of the GAAD study by fitting the competing methods (`skewBART`, BART, and SBART) to the mean PPD and mean CAL responses separately. All three Bayesian methods use 5,000 MCMC draws with 200 trees, and we compare their fits via the LPML. Results of the model comparison are summarized in Table 2.2. For both responses, the estimated LPMLs reveal that the `skewBART` outperforms the other two competing models substantially. The performance of SBART and BART are similar; SBART, which enables adaptation to the level of smoothness, provides only a slight boost compared to the BART. Also, the posterior estimates (95% credible intervals) of λ for CAL and PPD are 1.151 (1.129, 1.184)and 0.795 (0.778 0.820) obtained from fitting `skewBART`. Since λ is reparameterized in terms of α and the credible intervals for λ do not contain 0, it implies substantial evidence of right-skewness for both responses. This reveals that the assumption of a marginal SN error distribution is more appropriate than a Gaussian distribution for each response.

Second, we apply the `multi-skewBART` to jointly model the two skewed responses, mean CAL and mean PPD. We compare the fit to multi-BART via LPML, using 5000 MCMC draws and 200

Table 2.2: GAAD data analysis: Model fit summary measures (LPML) obtained from fitting the `skewBART`, SBART and BART models separately to the (mean) PPD and (mean) CAL responses.

	`skewBART`	SBART	BART
CAL	-380.407	-439.512	-442.230
PPD	-285.453	-338.455	-343.450

trees. As expected, the `multi-skewBART` allows for varying skewness levels (of the two responses) as well as dependence on the multivariate error terms, and hence outperforms (LPML = -694 vs -793.6) the multi-BART fit for the joint model. We conclude that the multivariate SN assumption presents an improvement for the GAAD study.

Next, we examine how the use of a multivariate model impacts the prediction accuracy under the SN assumption. Although both `skewBART` and `multi-skewBART` assume SN errors, the `multi-skewBART` allows sharing of information across decision trees and captures within-subject association. We compare the `skewBART` and `multi-skewBART` models using root mean squared error (RMSE), defined as $\text{RMSE}^2 = \sum_i \{Y_i^\star - \widehat{f}(X_i^\star)\}^2$ where the $(Y_i^\star, X_i^\star)$'s denote held-out test data. To make a direct comparison, the responses are standardized and the results are averaged over 20 splits into training and testing sets. Results are presented in Table 2.3. Although `multi-skewBART` doesn't adapt to smoothness levels, it gives the best performance in terms of RMSE. This result implies that the association between mean CAL and mean PPD within each subject plays a more essential role than the smoothness levels for the GAAD study. Also, the `multi-skewBART` shares information across the different responses by using the same decision trees to generate predictions for the two responses. The central message conveyed here is that separate modeling of mean PPD and mean CAL responses which ignores dependence between responses and which fails to share information across the different responses compromises the prediction accuracy.

Table 2.3: GAAD data analysis: Root mean squared error (RMSE) computed over 20 replications for the responses (PPD and CAL) from the `skewBART` and `multi-skewBART` fits.

	`skewBART`	`multi-skewBART`
CAL	1.345	0.838
PPD	0.626	0.423

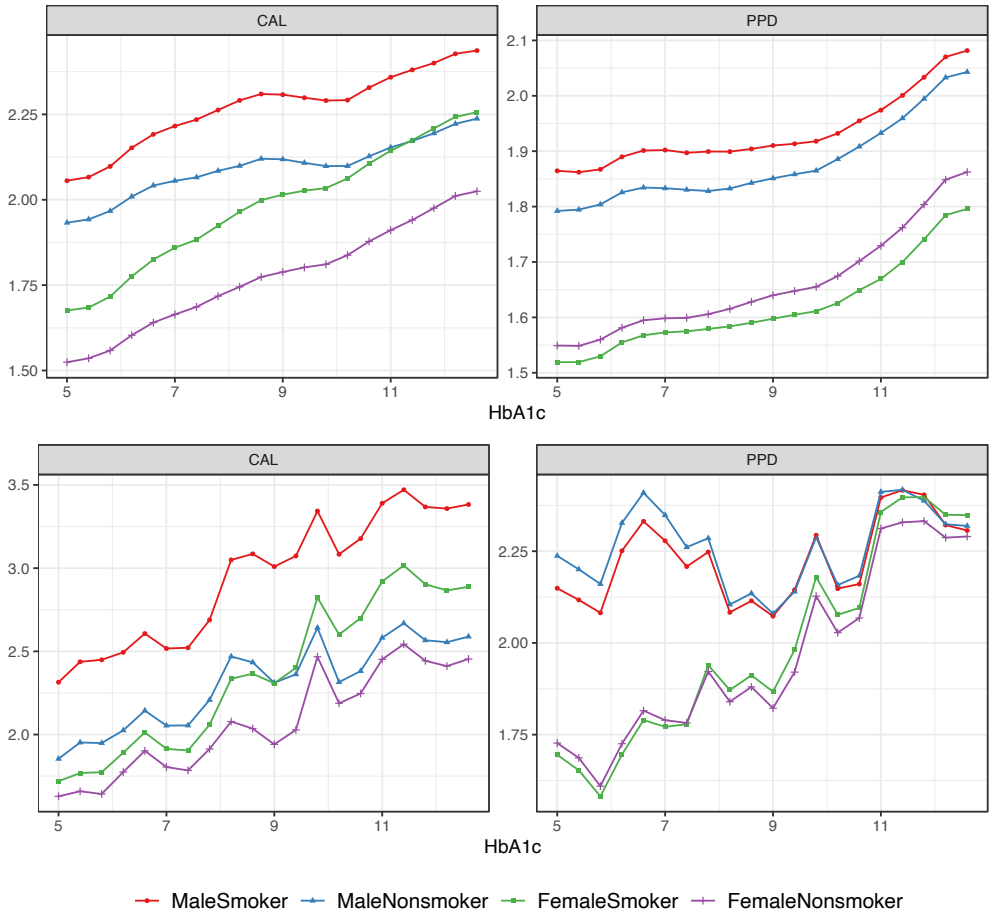

Figure 2.5: GAAD data analysis: Plots of the predicted CAL and PPD responses corresponding to the 4 subgroups (varying with gender and smoking status) from fitting the `skewBART` (upper panel), and `multi-skewBART` (bottom panel).

As there is existing strong evidence of a bi-directional association between Type-2 diabetes and PD[30, 21], we now focus on evaluating the association of varying HbA1c levels (an important biomarker of diabetes status) on the study responses (PPD and CAL). For subjects fixed at the median age (56 years) and BMI (33.987), we compare the posterior median predictions obtained from fitting `skewBART` and `multi-skewBART` models within the four subgroups defined by the combination of gender and smoking status. The results are displayed in Figure 2.5, with `skewBART` in the upper panel and `multi-skewBART` in the lower panel. We observe that with increasing HbA1c levels both responses exhibit an overall increasing trend, regardless of the subgroups or the model used. This reconfirms the (positive) association between HbA1c and PD. Interestingly, from the `skewBART` fit for

PPD, once gender is fixed, there does not exist any significant differences between the predictions for smokers and non-smokers. We also observe that the predictions from `skewBART` are smoother than the predictions from `multi-skewBART`, which exhibit large flucuations. Note that, while `skewBART` is capable of leveraging smoothness, it produces a higher RMSE on test data (see Table 2.3). This indicates that considering the association between responses for the GAAD data is more essential than trying to leverage smoothness in the regression function for obtaining good performance.

Finally, we examine whether the posterior median differences of the predicted PPD and CAL between male smokers and female nonsmokers are significant. From the `skewBART` fit, the 95% credible interval (CI) estimates of median differences of PPD and CAL lie above 0 over the HbA1c ranges (5.0, 11.8) and (5.0, 10.2), respectively. The corresponding HbA1c ranges from the `multi-skewBART` fit are $(5.8, 11.8)$ and $(6.2, 6.6)$, respectively. Thus the median differences remain > 0 over most of the range of HbA1c, pointing to the positive association observed between HbA1c and the PD responses.

2.5 Discussion

In this dissertation we proposed the `skewBART` model, and extended it to the `multi-skewBART` model which handles multivariate outcomes. The main idea is to use either a univariate SN density or an MSN density as the error density within the BART model. We showed that, when the error distribution is skewed, `skewBART` and `multi-skewBART` result in better model fit than regular BART; also, `multi-skewBART` enjoys additional benefits for multivariate responses due to its ability to account for within-subject dependence of the outcomes and because it uses multivariate decision trees to share information across regression functions. We showed that `multi-skewBART` produces better model fits on the GAAD data than fitting the two outcomes seperately using `skewBART`. Code for implementing these models is available on `GitHub` at `https://github.com/Seungha-Um/skewBART`.

The current work considers two continuous responses, PPD and CAL. However, in practice, data responses (elements in the multivariate response vector) can be of mixed types, such as *binary* 'bleeding on probing' outcomes in PD modeling. In such situations, considering an elegant latent variable, or factor modeling, framework with BART specifications may be worthwhile. lso, the number of available (non-missing) site-level responses (within each subject) can be correlated with the

PPD, or CAL responses, leading to the informative cluster size (ICS) scenario[25], and exploration of BART under ICS is non-existent. Computationally, we have relied on the Gibbs sampler and Metropolis-Hastings for tree updates. In big data settings (such as in observational PD databases), scalable Bayesian methods will likely be required. All of these are important avenues for future work, and will be considered elsewhere.

CHAPTER 3

BAYESIAN ADDITIVE REGRESSION TREES MODEL FOR MIXED RESPONSES

The presented methodology is motivated by a clinical study conducted at the Medical University of South Carolina (MUSC) to determine the periodontal disease status for Type-2 diabetic Gullah-speaking African-Americans [15]. The objective of this analysis is study the associations between disease status and patient-level covariates such as age, BMI, gender, HbAlC and smoking status. To measure periodontal status, hygienists often measure three of the most popular makers; clinical attachment loss (CAL), periodontal pocket depth(PPD) and bleeding on probing (BOP). It is common to summarize disease status using the whole-mouth mean CAL, PPD and the number of BOP sites for each subject above a certain threshold. Our data are a mixture of continuous and count responses, making joint modeling difficult. Using the Pólya-Gamma data augmentation technique, the multivariate framework can jointly model the continuous and count responses. Our proposed multivariate model can jointly analyze periodontal data from multiple measurements to improve estimation of disease status, and hence develop a more powerful method for studying the association between patient-level covariates and periodontal disease. We focus on the fact the same unknown set of features are associated with the responses and the mixed responses are correlated. In this dissertation, we present a methodology which allows for information to be shared nonparametrically for jointly modelling count and continuous responses using Bayesian sum-of-tree models.

3.1 The BART model for mixed responses

We assume our mixed responses has J types of responses. For the periodontal data, the $J = 3$ responses are meanCAL, meanPPD and the number of BOP sites for each patient. Note that mean CAL and mean PPD are continuous responses and the number of BOP sites is count response.

3.1.1 cBART: count response

We first introduce the cBART model, which is an extension of BART to accommodate count responses. We consider the negative binomial distribution with probability of success p_i and shape parameter r for the count response y_i^* for $i = 1, \ldots, N$. The effect of $\boldsymbol{X}_i$ is specified using a sum of trees $\sum_{t=1}^{m} g(\boldsymbol{X}_i, \mathcal{T}_t, \mathcal{M}_t)$ where $\mathcal{T}_t$ is a binary regression tree and $\mathcal{M}_t$ is a collection of leaf parameters for $\mathcal{T}$. We approximate the nonparametric regression function $f(\cdot)$ in (2.1) for the link function $\psi_i = log\dfrac{p_i}{1-p_i}$ as

$$\psi_i = \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{3.1}$$

For the cBART model, we use the SBART framework of Linero and Yang [29] to adapt to smoothness in $f(\cdot)$.

3.1.2 mBART: mixed response

We propose mBART which extends the sum-of-trees model (3.1) to accommodate mixed responses. This extension improves prediction accuracy by incorporating the correlation between continuous and count responses, as well as the dependency in error estimation.

Let y_{ij} and $y_{ij'}^*$ be the continuous and count response respectively for patient i for $i = 1, \ldots, N$. Note that $j = 1, \ldots, k$ and $j' = k+1, \ldots, J$ where k is the number of continuous responses and $(J-k)$ is the number of count responses. Like cBART, we assume $y_{ij}^*(s) \sim \mathcal{NB}(r, p_{ij})$ where the link function $\psi_{ij} = log\dfrac{p_{ij}}{1-p_{ij}}$. Define $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{ik})$ and $\boldsymbol{\psi}_i = (\psi_{ik+1}, \cdots, \psi_{iJ})$. Then, a new responses vector $\boldsymbol{L}_i = \{\boldsymbol{Y}_i, \boldsymbol{\psi}_i\}$ can be defined. Our proposed mBART, approximates the newly defined responses $\boldsymbol{L}_i$ as

$$\boldsymbol{L}_i = \sum_{t=1}^{T} \boldsymbol{g}_t^*(x_i, \mathcal{T}_t, \boldsymbol{M}_t) + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{MVN}(0, \Sigma)$$

where $\boldsymbol{g}_t^*(\boldsymbol{x}_i; \mathcal{T}_t, \boldsymbol{M}_t)$ returns a J-dimensional $\boldsymbol{\mu}_{t\ell}$ if $\boldsymbol{x}_i$ is associated with the leaf node ℓ in $\mathcal{T}_t$ for $\ell = 1, \ldots, n_t$. The tree $\mathcal{T}_t$ in (2.4) has a binary tree structure like univariate cBART but $\boldsymbol{M}_t = \{\boldsymbol{\mu}_{t1}, \cdots, \boldsymbol{\mu}_{tn_t}\}$ is now a set of $J-$dimensional leaf parameters. The joint model with the mixed responses improves prediction accuracy by incorporating the correlation between responses, as well as the dependency in error estimation.

3.2 Prior specification and posterior computation

In the literature it has been seen empirically that the default priors proposed by Chipman et al. [12] work remarkably well in practice. For the tree structures, we use the default prior for the BART model and provide priors for rest of the parameters for the negative binomial regression. We then derive Gibbs samplers for fitting the models.

3.2.1 `cBART`: prior choices and MCMC computation

Each column of the design matrix X is distributed approximately uniformly on $[0, 1]$ by independent quantile normalizations. Additionally, following Chipman et al.[12], the dependent variable Y is standardized by first shifting and rescaling. Then, the leaf parameters $\mu_{t\ell}$ are assumed independent and identically distributed as $\mathcal{N}(0, \sigma_\mu^2/m)$. With this prior, the leaf parameters $\mu_{t\ell}$ are shrunk towards zero, thus regularizing the effect of the individual tree components to contribute only a small part in the overall fit.

For choosing the bandwidth κ_b, Linero and Yang [29] recommend using tree specific κ_t's shared across branches in a fixed tree, with $\kappa_t \sim \text{Exp}(0.1)$. We specify a half-Cauchy prior $\sigma \sim \text{Cauchy}_+(0, \hat{\tau})$, where $\hat{\sigma}$, where $\hat{\sigma}$ is chosen empirically. We obtain $\hat{\sigma}$ by fitting the linear regression using the log transformation.

To sample ψ_i using the Polya-Gamma Distribution, denote ω_i as a random variable drawn from the Polya-Gamma (PG) distribution as

$$\omega_i \sim PG\left(y_i + r, 0\right)$$

We have $\mathbb{E}_{\omega_i}\left[\exp\left(-\omega_i\psi_i^2/2\right)\right] = \cosh^{-(y_i+r)}\left(\psi_i/2\right)$. Thus the likelihood of ψ_i can be expressed as

$$\begin{aligned}\mathscr{L}\left(\psi_i\right) &\propto \frac{\left(e^{\psi_i}\right)^{y_i}}{\left(1+e^{\psi_i}\right)^{y_i+r}} \\ &\propto \exp\left(\frac{y_i - r}{2}\psi_i\right)\mathbb{E}_{\omega_i}\left[\exp\left(-\omega_i\psi_i^2/2\right)\right]\end{aligned}$$

Given the values of $\{\omega_i\}$ and the PG prior, the conditional posterior of ψ can be expressed as

$$\left(\boldsymbol{\psi} \mid -\right) \propto \mathscr{N}\left(\boldsymbol{\psi}; \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t), \sigma^2\mathbf{I}\right)\prod_{i=1}^{N} e^{-\frac{\omega_i}{2}}\left(\psi_i - \frac{y_i - r}{2\omega_i}\right)^2$$

Thus, the closed-form updates of ψ_i is obtained by exploiting conditional conjugacy via a Polya-Gamma distribution based data augmentation approach. Also, given the values of $\boldsymbol{\psi}$ and the PG prior, the conditional posterior of ω_i is computed as

$$(\omega_i \mid -) \propto \exp\left(-\omega_i \psi_i^2/2\right) PG\left(\omega_i; y_i + r, 0\right). \tag{3.2}$$

In summary, our proposed prior is given by

$$\begin{aligned}
&y_i \sim \text{NB}(r, p_i) && p_i = \frac{e^{\psi_i}}{1+e^{\psi_i}} \\
&\psi_i = \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \epsilon_i, && r \sim \text{Gamma}(a, b) \\
&\epsilon_i \sim \mathcal{N}(0, \sigma^2), && \sigma \sim \text{Cauchy}_+(0, \hat{\sigma}) \\
&\alpha = 0.95, && \beta = 2.
\end{aligned}$$

Below, we describe our MCMC algorithm for fitting `mBART` via data augmentation. Our strategy is to augment the latent variables ψ_i within the Bayesian backfitting algorithm[20] in order to sample approximately from the posterior distribution

$$\pi((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \sigma^2, \mid \boldsymbol{\psi})$$

using MCMCs, where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$. By abuse of notation, we use the symbol $[\theta \mid \gamma]$ to denote the conditional posterior distribution of θ given γ. Our Gibbs sampler draws $(\mathcal{T}_j, \mathcal{M}_j)$ for $j = 1, \ldots, m$ given $(\mathcal{T}_{-t}, \mathcal{M}_{-t}, \sigma^2)$ from

$$[\mathcal{T}_j, \mathcal{M}_j | \mathcal{T}_{-j}, \mathcal{M}_{-j}, \boldsymbol{Z}, \boldsymbol{\psi}, \sigma^2], \tag{3.3}$$

where $\mathcal{M}_{-j} = \{\mathcal{M}_t : t \neq j\}$ and $\mathcal{T}_{-j} = \{\mathcal{T}_t : t \neq j\}$. Then, samples of σ^2 is drawn from conditional posterior distributions

$$[\sigma^2 | \mathcal{T}_1, \ldots, \mathcal{T}_m, \mathcal{M}_1, \ldots, \mathcal{M}_m, \boldsymbol{\psi}].$$

To implement the m draws of $(\mathcal{T}_j, \mathcal{M}_j)$ in (3.3), the draws are equivalent to sampling from

$$[\mathcal{T}_j, \mathcal{M}_j | \boldsymbol{R}_j, \sigma^2] \ , \tag{3.4}$$

where $\boldsymbol{R}_j = \boldsymbol{\psi} - \sum_{k \neq j} g(X; \mathcal{T}_k, \mathcal{M}_k)$. The Bayesian backfitting algorithm for `skewBART` is described in the Algorithm 3 below, with the exact conditional posterior distributions relegated to the Appendix.

Algorithm 2 Bayesian backfitting algorithm(cBART)

1: **for** $t = 1$ to m **do**
2: Set $\boldsymbol{R}_t \leftarrow \boldsymbol{\psi} - \sum_{k \neq t} g(X; \mathcal{T}_k, \mathcal{M}_k)$
3: Propose a new tree $\mathcal{T}_t^* \sim Q(\mathcal{T}_t \rightarrow \mathcal{T}_t^*)$
4: Sample $U \sim Uniform(0, 1)$ and compute the acceptance ratio

$$\alpha = \frac{\Lambda(\mathcal{T}_t^*)Q(\mathcal{T}_t^* \rightarrow \mathcal{T}_t)}{\Lambda(\mathcal{T}_t)Q(\mathcal{T} \rightarrow \mathcal{T}_t^*)}$$

5: Accept a draw if $U \leq \alpha$, otherwise reject a draw.
6: Update the leaf node parameters.
7: **end for**
8: **for** i=1 to N **do**
9: Sample ψ_i from $\psi_i \sim \mathcal{N}(\mu_\psi, \Sigma_\psi)$.
10: Sample ω_i from $\omega_i \sim PG(y_i + r, \psi_i)$.
11: **end for**
12: Sample r from M-H.

3.2.2 `mBART`: prior choices and MCMC computation

To extend the `cBART` model to handle mixed J-dimensional responses, we need to replace the univariate priors with priors on vector/matrix parameters. With the newly defined responses $\boldsymbol{L}_i = \{\boldsymbol{Y}_i, \boldsymbol{\psi}_i\}$, we use the conjugate multivariate normal distribution, $\mathcal{N}_k(\boldsymbol{\mu}_M, \Sigma_M)$ for $\pi(\boldsymbol{\mu}_{t\ell} \mid \mathcal{T}_t)$, that allows for (1.3) to be computed in closed form. The hyper parameters $\boldsymbol{\mu}_M$ and Σ_M are chosen to assign high probability to the range $(\boldsymbol{L}_{\min}, \boldsymbol{L}_{\max})$, where the j'th row of this matrix is given by $(\min_i Y_{ij}, \max_i Y_{ij})$ for continuous response and $(\min_i \psi_{ij}, \max_i \psi_{ij})$ for count response. Note that the range $(\min_i \psi_{ij}, \max_i \psi_{ij})$ can be estimated with the `cBART` model. We assign an inverse-Wishart prior distribution to Σ.

Like `cBART`, the closed-form updates of ψ_{ij} is obtained by exploiting conditional conjugacy via a Polya-Gamma distribution based data augmentation approach and given the values of $\boldsymbol{\psi}$ and the

PG prior, the conditional posterior of ω_i can be obtained. We summarize our prior choices below:

$$
\begin{aligned}
&y^*_{ij} \sim \text{NB}(r, p_{ij}) && p_{ij} = \frac{e^{\psi_{ij}}}{1+e^{\psi_{ij}}} \\
&(\boldsymbol{Y}_i, \boldsymbol{\psi}_i)' = \sum_{t=1}^{m} \boldsymbol{g}^*_t(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t) + \boldsymbol{\epsilon}_i && \boldsymbol{\epsilon}_i \sim \text{MVN}_k(0, \Sigma), \\
&r \sim \text{Gamma}(a, b) && \Sigma \sim \text{inverse-Wishart}(\nu_0, S_0^{-1}), \\
&\alpha = 0.95, && \beta = 2.
\end{aligned}
$$

The Gibbs sampler draws $(\mathcal{T}_j, \boldsymbol{M}_j)$ given $(\mathcal{T}_{-t}, \boldsymbol{M}_{-t}, \Sigma)$ for $j = 1, \ldots, m$ from

$$[\mathcal{T}_j, \boldsymbol{M}_j | \tilde{\boldsymbol{R}}_j, \Sigma] \ ,$$

where $\tilde{\boldsymbol{R}}_j = \boldsymbol{L} - \sum_{k \neq j} \boldsymbol{g}^*(X; \mathcal{T}_k, \boldsymbol{M}_k)$. Then, we sample of Σ from each full conditional distribution given as

$$[\Sigma | \mathcal{T}_1, \ldots, \mathcal{T}_m, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_m, \boldsymbol{L}].$$

The MCMC algorithm for this model is similar to the original MCMC scheme for the `cBART`, but now extended to the multivariate setting.

3.3 Future work

In this dissertation, we proposed the cBART, and extended it to the mBART model which handles mixed responses. The main idea is to use the Pólya-Gamma data augmentation technique to jointly model the continuous and count responses. Our proposed multivariate model can jointly analyze periodontal data from multiple measurements to improve estimation of disease status, and hence develop a more powerful method for studying the association between patient-level covariates and periodontal disease. Also, using multivariate decision trees to share information across regression functions, mBART can lead to improvements in fit compared to the competing methods. For the future work, we hope to show that mBART produces better model fits on the GAAD data than fitting the mixed responses separately.

APPENDIX A

SUPPLEMENTAL MATERIAL

A.1 Proof of equation (1.7)

LEMMA Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Theh for any fixed k -dimensional vector $\mathbf{a}$ and $k \times n$ matrix $\mathbf{B}$,

$$E\left[\Phi_k(\mathbf{a} + \mathbf{B}\mathbf{Y} \mid \eta, \boldsymbol{\Omega})\right] = \Phi_k\left(\mathbf{a} \mid \boldsymbol{\eta} - \mathbf{B}\boldsymbol{\mu}, \boldsymbol{\Omega} + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T\right)$$

Let $\mathbf{Y} = \boldsymbol{\Delta}\,|\mathbf{Z}| + \mathbf{U}$. Since $\mathbf{Y} \mid \mathbf{T} = \mathbf{t} \sim N_n(\boldsymbol{\Delta}\mathbf{t} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{T} = |\mathbf{Z}| \sim HN_n(\mathbf{0}, \mathbf{I}_n)$ (the multivariate standardized half-normal distribution), then by **LEMMA** and the fact that $\left(\mathbf{I}_n + \Delta\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Delta}\left(\boldsymbol{\Sigma} + \boldsymbol{\Delta}^2\right)^{-1}$ it follows that

$$\begin{aligned}
f_{\mathbf{Y}}(\mathbf{w}) =& \int_{\mathbb{R}_+^n} \phi_n(\mathbf{w} \mid \boldsymbol{\Delta}\mathbf{t} + \boldsymbol{\mu}, \boldsymbol{\Sigma}) 2^n \phi_n(\mathbf{t}) d\mathbf{t} \\
=& 2^n \int_{\mathbb{R}_+^n} \phi_n\left(\mathbf{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Delta}^2\right) \\
& \times \phi_n\left(\mathbf{t} \mid \left(\mathbf{I}_n + \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}), \left(\mathbf{I}_n + \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\right)^{-1}\right) d\mathbf{t} \\
=& 2^n \phi_n\left(\mathbf{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Delta}^2\right) \Phi_n\left(\boldsymbol{\Delta}\left(\boldsymbol{\Sigma} + \boldsymbol{\Delta}^2\right)^{-1}(\mathbf{w} - \boldsymbol{\mu}) \mid \mathbf{0}, \left(\mathbf{I}_n + \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\right)^{-1}\right),
\end{aligned}$$

i.e. $\mathbf{Y} \overset{d}{=} \mathbf{Y} \sim SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$, which concludes the proof.

A.2 Proof of equation (1.8)

To prove

$$\frac{\left(e^{\psi}\right)^a}{\left(1 + e^{\psi}\right)^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) \mathrm{d}\omega$$

where ω is a PG random variable, plug $a = \kappa + b/2$,

$$\frac{\left(e^{\psi}\right)^{a}}{\left(1+e^{\psi}\right)^{b}} = \frac{\left(e^{\psi}\right)^{\kappa+b/2}}{\left(1+e^{\psi}\right)^{b}} \tag{A.1}$$

$$= \frac{\left(e^{\psi}\right)^{\kappa}\left(e^{\psi}\right)^{b/2}}{\left(1+e^{\psi}\right)^{b}} \tag{A.2}$$

$$= \frac{\left(e^{\psi}\right)^{\kappa}}{\left(1+e^{\psi}\right)^{b}\left(e^{-\psi/2}\right)^{b}} \tag{A.3}$$

$$= \frac{\left(e^{\psi}\right)^{\kappa}}{\left(\frac{1+e^{\psi}}{e^{\psi/2}}\right)^{b}} \tag{A.4}$$

Since $\cosh(x) = \frac{e^{x}+e^{-x}}{2}$, we have

$$\begin{aligned}\left(\frac{1+e^{\psi}}{e^{\psi/2}}\right)^{b} &= \left(\frac{e^{\psi}}{e^{\psi/2}} + \frac{1}{e^{\psi/2}}\right)^{b} \\ &= \left(e^{\psi/2} - e^{-\psi/2}\right)^{2} \\ &= \left[2\cosh(\psi/2)\right]^{b}.\end{aligned}$$

Then, the equation of A.1 can be rewritten as

$$\frac{\left(e^{\psi}\right)^{a}}{\left(1+e^{\psi}\right)^{b}} = \frac{\left(e^{\psi}\right)^{\kappa}}{\left(\frac{1+e^{\psi}}{e^{b/2}}\right)^{b}} = \frac{\left(e^{\psi}\right)^{\kappa}}{\left[2\cosh(\psi/2)\right]^{b}} = \frac{2^{-b}\left(e^{\psi}\right)^{\kappa}}{\cosh^{n}(\psi/2)} \tag{A.5}$$

By logit transformation,

$$\mathbb{E}[\exp(-\omega t)] = \frac{1}{\cosh^{b}(\sqrt{t/2})}$$

setting $t = \psi^{2}/2$:

$$\mathbb{E}\left[\exp\left(-\omega\frac{\psi^{2}}{2}\right)\right] = \frac{1}{\cosh^{b}(\psi/2)} \tag{A.6}$$

Finally, plug (A.6) into (A.5), and we're done:

$$\frac{\left(e^{\psi}\right)^{a}}{\left(1+e^{\psi}\right)^{b}} = \frac{2^{-b}\left(e^{\psi}\right)^{\kappa}}{\cosh^{n}(\psi/2)} = 2^{-b}e^{\psi\kappa}\mathbb{E}\left[\exp\left(-\omega\frac{\psi^{2}}{2}\right)\right]$$

.

A.3 MCMC for the `skewBART` fitting

We provide details on the Bayesian backfitting algorithm for the `skewBART` model. Recall that the model is given by

$$
\begin{aligned}
Y_i &= \sum_{t=1}^{m} g(\boldsymbol{x}_i; \mathcal{T}_t \mathcal{M}_t) + \lambda |Z_i| + W_i, \qquad W_i \overset{iid}{\sim} \mathcal{N}(0, \tau), \\
\sqrt{\tau} &\sim Cauchy_+(0, \hat{\tau}), \\
\lambda &\sim \mathcal{N}(0, \delta), \\
Z_i &\sim \mathcal{N}(0, 1).
\end{aligned}
$$

Define $R_i = Y_i - f(X_i) = \lambda |Z|_i + \epsilon_i$, then $[R_i | \lambda, |Z|_i] \sim \mathcal{N}(\lambda Z_i, \tau)$. The posterior distribution of $|Z_i|$ is computed as

$$
\begin{aligned}
p(Z_i | \lambda, R_i) &\propto p(R_i | \lambda, Z_i) p(Z_i) \\
&\propto \exp\bigg(- \frac{(R_i - \lambda Z_i)^2}{2\tau} \bigg) \exp\bigg(- \frac{Z_i^2}{2} \bigg) \\
&\propto \exp\bigg[- \frac{1}{2\tau} \bigg\{ (\lambda^2 + \tau) Z_i^2 - 2\lambda R_i Z_i \bigg\} \bigg] \\
&\propto \exp\bigg[- \frac{\lambda^2 + \tau}{2\tau} \bigg\{ Z_i^2 - 2 \frac{\lambda R_i}{\lambda^2 + \tau} Z_i \bigg\} \bigg]
\end{aligned}
$$

Thus, $|Z_i| | \lambda, R_i \sim \mathcal{N}_+\bigg(\frac{\lambda R_i}{\lambda^2 + \tau}, \frac{\tau}{\lambda^2 + \tau} \bigg)$.

Define $\boldsymbol{R} = (R_1, \cdots, R_n)$ and $\boldsymbol{Z} = (|Z_1|, \cdots, |Z_n|)$. Then,

$$
\begin{aligned}
p(\lambda | \tau, \mathbf{R}, \mathbf{Z}) &\propto p(\mathbf{R} | \lambda, \tau, \mathbf{Z}) p(\lambda) \\
&\propto \exp\bigg[- \frac{1}{2\tau} (\mathbf{R} - \lambda \mathbf{Z})^T (\mathbf{R} - \lambda \mathbf{Z}) \bigg] exp\bigg[- \frac{1}{2\delta} \lambda^2 \bigg] \\
&\propto \exp\bigg[- \frac{1}{2} (\mathbf{Z}^T \mathbf{Z}/\tau + 1/\delta) \bigg(\lambda^2 - 2 \frac{\mathbf{Z}^T \mathbf{R}/\tau}{\mathbf{Z}^T \mathbf{Z}/\tau + 1/\delta} \lambda \bigg) \bigg]
\end{aligned}
$$

Thus, the update of λ is $\lambda | \mathbf{R}, \mathbf{Z} \sim \mathcal{N}\bigg(\frac{\mathbf{Z}^T \mathbf{R}}{\mathbf{Z}^T \mathbf{Z} + \tau/\delta}, (\mathbf{Z}^T \mathbf{Z}/\tau + 1/\delta)^{-1} \bigg)$

To update τ, under a flat prior for $1/\tau$, the full-conditional of $1/\tau$ is $Gamma(n/2 + 1, ||R^*||^2)$ where $R^* = (\mathbf{R} - \lambda \mathbf{Z})$. We assume $\sqrt{\tau} \sim Cauchy_+(0, \sqrt{\hat{\tau}}_{lasso})$We use this full-conditional under the

flat prior as a proposal distribution for $1/\tau$, after adjusting for the Jacobian of the transformation, the acceptance probability becomes

$$\begin{aligned} A(\tau \to \tau') &= \frac{p(\tau'|R^*)q(\tau|\tau')}{p(\tau|R^*)q(\tau'|\tau)} \wedge 1 \\ &= \frac{p(\tau')p(R^*|\tau')q(\tau|\tau')}{p(\tau)p(R^*|\tau)q(\tau'|\tau)} \wedge 1 \\ &= \frac{Cauchy_+(\tau'|0, \hat{\tau}_{lasso})\tau'^3}{Cauchy_+(\tau|0, \hat{\tau}_{lasso})\tau^3} \end{aligned}$$

A.4 MCMC for the `multi-skewBART` fitting

We provide details on the Bayesian backfitting algorithm for the `multi-skewBART` model. Recall that the hierarchical model for `multi-skewBART` is

$$\begin{aligned} \boldsymbol{Y}_i &= \sum_{t=1}^{T} \boldsymbol{g}_t^*(\boldsymbol{x}_i, \mathcal{T}_t, \boldsymbol{M}_t) + \Delta|\boldsymbol{Z}_i| + \boldsymbol{W}_i, \qquad \boldsymbol{W}_i \overset{iid}{\sim} \mathcal{N}_k(0, \Sigma) \\ \boldsymbol{Z}_i &\sim \mathcal{N}(0, \boldsymbol{I}_k), \\ \boldsymbol{\lambda} &\sim \mathcal{N}_k(0, \Psi) \\ \Sigma &\sim inverse\text{-}Wishart(\nu_0, S_0^{-1}) \end{aligned}$$

Define

$$\boldsymbol{R}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} - \begin{pmatrix} f_1(X_i) \\ f_2(X_i) \end{pmatrix} = \begin{pmatrix} \lambda_1 Z_i \\ \lambda_2 Z_i \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix}.$$

Then the posterior distribution of $|\boldsymbol{Z}_i|$ is computed as

$$\begin{aligned} p(\boldsymbol{Z}_i|\Delta, \boldsymbol{R}_i) &\propto p(\boldsymbol{R}_i|\Delta, \boldsymbol{Z}_i)p(\boldsymbol{Z}_i) \\ &\propto \exp\left[-\frac{1}{2}(\boldsymbol{R}_i - \Delta\boldsymbol{Z}_i)^T\Sigma^{-1}(\boldsymbol{R}_i - \Delta\boldsymbol{Z}_i)\right] \exp\left(-\frac{\boldsymbol{Z}_i^2}{2}\right) \\ &\propto \exp\left[\left\{-\frac{1}{2}\boldsymbol{Z}_i^T\Delta^T\Sigma^{-1}\Delta\boldsymbol{Z}_i - 2\boldsymbol{Z}_i^T\Delta^T\Sigma^{-1}\boldsymbol{R}_i + \boldsymbol{Z}_i^2\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\boldsymbol{Z}_i^T(\Delta^T\Sigma^{-1}\Delta + I)\boldsymbol{Z}_i - 2\boldsymbol{Z}_i^T\Delta^T\Sigma^{-1}\boldsymbol{R}_i\right\}\right] \end{aligned}$$

Therefore, $|\boldsymbol{Z}_i|$ is updated as

$$[|\boldsymbol{Z}_i| \mid \cdot\,] \sim \mathcal{N}_{k+}\left((\Delta\Sigma^{-1}\Delta + I_k)^{-1}\Delta\Sigma^{-1}\mathbf{R}_i^*,\ \Delta\Sigma^{-1}\Delta + I_k)^{-1}\right).$$

Here $\mathcal{N}_{k+}(\mu, \Sigma)$ denotes a k-dimensional normal distribution with mean μ and covariance Σ truncated to the region $\{\mathbf{z} \in \mathbb{R}^k : z_j \geq 0, j = 1, \ldots, k\}$.

The posterior distribution of $\boldsymbol{\lambda}$ is

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{Z}, \boldsymbol{R}, \Sigma) &\propto p(\boldsymbol{R}|\boldsymbol{\lambda}, \boldsymbol{Z}, \Sigma)p(\boldsymbol{\lambda}) \\ &\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{R}_i - Z_i\boldsymbol{\lambda})^T\Sigma^{-1}(\boldsymbol{R}_i - Z_i\boldsymbol{\lambda})\right]\exp\left[-\frac{1}{2}\boldsymbol{\lambda}^T\Lambda^{-1}\boldsymbol{\lambda}\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}Z_i^2\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{\lambda} - 2\sum_{i=1}^{n}Z_i\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{R}_i + \boldsymbol{\lambda}^T\Lambda^{-1}\boldsymbol{\lambda}\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\boldsymbol{\lambda}^T\left(\sum_{i=1}^{n}Z_i^2\Sigma^{-1} + \Lambda^{-1}\right)\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\Sigma^{-1}\sum_{i=1}^{n}Z_i\boldsymbol{R}_i\right\}\right] \end{aligned}$$

Thus $\Delta = \text{diag}(\boldsymbol{\lambda})$ is updated as

$$[\boldsymbol{\lambda}|\cdot\,] \sim \mathcal{N}_k\left(\left(\sum_{i=1}^{n}M_i\Sigma^{-1}M_i + \Psi^{-1}\right)^{-1}\sum_{i=1}^{n}M_i\Sigma^{-1}\mathbf{R}_i^*,\ \left(\sum_{i=1}^{n}M_i\Sigma^{-1}M_i + \Psi^{-1}\right)^{-1}\right)$$

where $M_i = diag(Z_{i1}, \cdots, Z_{ik})$.

The update of Σ, recall that $\Sigma \sim inverse\text{-}Wishart(\nu_0, S_0^{-1})$ such that

$$p(\Sigma) = \left[2^{\nu_0 p/2}\pi^{\binom{p}{2}/2}|S_0|^{-\nu_0/2}\prod_{j=1}^{p}\Gamma([\nu_0 + 1 - j]/2)\right]^{-1}|\Sigma|^{-(\nu_0+p+1)/2}\exp\left\{-\frac{1}{2}tr(S_0\Sigma^{-1})\right\}$$

If we are confident that the true covariance matrix is near some covariance matrix Σ_0, then we might choose ν_0 to be large and set $S_0 = (\nu_0 - p - 1)\Sigma_0$, making the distribution of Σ concentrated around Σ_0. On the other hand, choosing $\nu_0 = p + 2$ and $S_0 = \Sigma_0$ makes Σ only loosely centered aground Σ_0. Thus, the posterior distribution of Σ is

$$\begin{aligned} p(\Sigma|\boldsymbol{\lambda}, \boldsymbol{R}, \boldsymbol{Z}) &\propto p(\boldsymbol{R}|\sigma, \boldsymbol{\lambda}, \boldsymbol{Z})p(\Sigma) \\ &\propto \left(\frac{1}{2\pi|\Sigma|^{1/2}}\right)^n\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{R}_i - Z_i\boldsymbol{\lambda})^T\Sigma^{-1}(\boldsymbol{R}_i - Z_i\boldsymbol{\lambda})\right]\Sigma^{-(\nu_0+2+1)/2}\exp\left[-\frac{1}{2}tr(S_0\Sigma^{-1})\right] \end{aligned}$$

Note that $\sum_{i=1}^{n} b_i^T A b_i = tr(BAB^T) = tr(B^T BA)$ where B is a matrix whose k^{th} row is b_k^T. In this sense, $\sum_{i=1}^{n}(\boldsymbol{R}_i - Z_i\boldsymbol{\lambda})^T\Sigma^{-1}(\boldsymbol{R}_i - Z_i\boldsymbol{\lambda}) = tr(B^T B\Sigma^{-1})$ where B is a matrix whose k^{th} row is $(\boldsymbol{R}_i - Z_i\boldsymbol{\lambda})^T$. Thus, the posterior distribution is rewritten as

$$\begin{aligned} p(\Sigma|\boldsymbol{\lambda}, \boldsymbol{R}, \boldsymbol{Z})] &\propto |\Sigma|^{-n/2} \exp\left[-\frac{1}{2} tr(B^T B\Sigma^{-1})\right] |\Sigma|^{-(\nu_0+2+1)/2} \exp\left[-\frac{1}{2} tr(S_0\Sigma^{-1})\right] \\ &\propto |\Sigma|^{-(\nu_0+n+2+1)/2} \exp\left[-\frac{1}{2} tr((B^T B + S_0)\Sigma^{-1})\right] \end{aligned}$$

The Σ is updated as

$$\Sigma \sim \text{inverse-Wishart}\left(\nu_0 + n, (B^T B + S_0)^{-1}\right)$$

where B is a matrix whose kth row is $(\boldsymbol{R}_i^* - \boldsymbol{Z}_i\boldsymbol{\lambda})^T$. The Bayesian backfitting algorithm for fitting `multi-skewBART` is now essentially the same as the algorithm for `skewBART` but with the updates of $(\Sigma, \{\boldsymbol{Z}_i\}, \boldsymbol{\lambda})$ replacing the updates for $(\tau, \{Z_i\}, \lambda)$.

Algorithm 3 Bayesian backfitting algorithm(`multi-skewBART`)

1: **for** $t = 1$ to m **do**
2: Set $\boldsymbol{R}_t \leftarrow \boldsymbol{Y} - \sum_{k\neq t} g(X; \mathcal{T}_k, \mathcal{M}_k) - \lambda\boldsymbol{Z}$
3: Propose a new tree $\mathcal{T}_t^* \sim Q(\mathcal{T}_t \to \mathcal{T}_t^*)$
4: Compute the acceptance ratio

$$r = \frac{\Lambda(\mathcal{T}_t^*)Q(\mathcal{T}_t^* \to \mathcal{T}_t)}{\Lambda(\mathcal{T}_t)Q(\mathcal{T} \to \mathcal{T}_t^*)}$$

5: Accept a draw if $U \leq r$ where $U \sim Uniform(0, 1)$, otherwise reject a draw.
6: Update the leaf node parameters.
7: **end for**
8: **for** i=1 to N **do**
9: Sample $\boldsymbol{Z}_i$ from the full conditional distribution of $\boldsymbol{Z}_i$.
10: **end for**
11: Sample Σ from the full conditional distribution of Σ.
12: Update Δ from the full conditional distribution of Δ.

A.5 The integrated likelihood for `multi-skewBART`

To update $\mathcal{T}$, the conditional posterior of $\mathcal{T}_t$ given the rest of the parameters is required. The derivation of this conditional posterior for `multi-skewBART` is actually a generalization of the derivation of the conditional posterior of $\mathcal{T}_t$ for univariate `skewBART`, and we provide only the sketch of the derivation of this conditional posterior for `multi-skewBART`. Define $\boldsymbol{S}_i = \boldsymbol{Y}_i - \Delta \boldsymbol{Z}_i$ and $\boldsymbol{\eta}_i = \mathbf{S}_i - \sum_{j \neq t} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M})$. Let $\mathcal{L}_t$ denote the collection of leaf nodes of tree t and let $[x \rightsquigarrow \ell]$ mean that the covariate value x is associated to leaf node ℓ of tree t. The conditional posterior is

$$\begin{aligned} p(\mathcal{T}_t \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}) &= \pi_{\mathcal{T}}(\mathcal{T}_t) \int \prod_{i=1}^{n} \phi_k(\mathbf{S}_i | \mathbf{f}(\mathbf{X}_i), \Sigma) \left[\prod_{\ell \in \mathcal{L}_t} \phi_k(\boldsymbol{\mu}_{tl} | 0, D) \right] d\boldsymbol{\mu}_{t\ell} \\ &= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} \int \prod_{i: \mathbf{X}_i \rightsquigarrow \ell} \{ \phi_k(\boldsymbol{\eta}_i | \boldsymbol{\mu}_{t\ell}, \Sigma) \} \phi_k(\boldsymbol{\mu}_{t\ell} | 0, D) d\boldsymbol{\mu}_{t\ell}, \end{aligned}$$

where $\phi_m(U | \mu, \Sigma)$ is the multivariate $\mathcal{N}_m(\mu, \Sigma)$ density and $\boldsymbol{\mu}_{t\ell}$ has $\mathcal{N}_k(\mathbf{0}, D)$ prior. This integrated likelihood can be computed in closed form easily by using standard properties of the multivariate Gaussian distribution.

Define $\boldsymbol{S}_i = \boldsymbol{Y}_i - \Delta \boldsymbol{Z}_i$ and $\boldsymbol{\eta}_i = \mathbf{S}_i - \sum_{j \neq t} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M})$. Assume $\boldsymbol{\mu}_{t\ell} \sim MNV(\mathbf{0}, D)$.

$$\Lambda(\mathcal{T}) = \pi_{\mathcal{T}}(\mathcal{T}_t) \int \prod_{i=1}^{n} p(\mathbf{S}_i|\mathbf{f}(\mathbf{X}_i), \boldsymbol{\omega}) \left[\prod_{\ell \in \mathcal{L}_t} \pi_{\boldsymbol{\mu}}(\boldsymbol{\mu}_{tl}) \right] d\boldsymbol{\mu}_{tl}$$

$$= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} \int \prod_{i:\mathbf{X}_i \rightsquigarrow (t,\ell)} MVN\left(\mathbf{S}_i|\mathbf{f}(\mathbf{X}_i), \Sigma\right) MVN(\boldsymbol{\mu}_{t\ell}) d\boldsymbol{\mu}_{t\ell}$$

$$= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} \int \prod_{i:\mathbf{X}_i \rightsquigarrow (t,\ell)} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{S}_i - \mathbf{f}(\mathbf{X}_i))^T \Sigma^{-1} (\mathbf{S}_i - \mathbf{f}(\mathbf{X}_i)) \right]$$

$$\times (2\pi)^{-p/2} |D|^{-1/2} \exp\left[-\frac{1}{2} \boldsymbol{\mu}_{t\ell}^T D^{-1} \boldsymbol{\mu}_{t\ell} \right] d\boldsymbol{\mu}_{tl}$$

$$= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} \int (2\pi)^{-pN_\ell/2} |\Sigma|^{-N_\ell/2} (2\pi)^{-p/2} |D|^{-1/2} \exp\left[-\frac{1}{2} \boldsymbol{\mu}_{t\ell}^T D^{-1/2} \boldsymbol{\mu}_{t\ell} \right]$$

$$\times \exp\left[-\frac{1}{2} \sum_{i:\mathbf{X}_i \rightsquigarrow (t,\ell)} \left(\mathbf{S}_i - \sum_{j \neq t} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M}) - \boldsymbol{\mu}_{t\ell} \right)^T \Sigma^{-1} \left(\mathbf{S}_i - \sum_{j \neq t} g(\mathbf{X}_i; \mathcal{T}_t, \mathcal{M}) - \boldsymbol{\mu}_{t\ell} \right) \right] d\boldsymbol{\mu}_{tl}$$

$$= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} \int (2\pi)^{-pN_\ell/2} |\Sigma|^{-N_\ell/2} (2\pi)^{-p/2} |D|^{-1/2} \exp\left[-\frac{1}{2} \boldsymbol{\mu}_{t\ell}^T D^{-1} \boldsymbol{\mu}_{t\ell} \right]$$

$$\times \exp\left[-\frac{1}{2} \left\{ N_\ell \boldsymbol{\mu}_{t\ell}^T \Sigma^{-1} \boldsymbol{\mu}_{t\ell} - 2\boldsymbol{\mu}_{t\ell}^T \Sigma^{-1} \sum_i \boldsymbol{\eta}_i + \sum_i \boldsymbol{\eta}_i^T \Sigma^{-1} \boldsymbol{\eta}_i \right\} \right] d\boldsymbol{\mu}_{tl}$$

$$= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} (2\pi)^{-pN_\ell/2} |\Sigma|^{-N_\ell/2} (2\pi)^{-p/2} |D|^{-1/2}$$

$$\int \exp\left[-\frac{1}{2} \left(\boldsymbol{\mu}_{t\ell}^T \left(D^{-1} + N_\ell \Sigma^{-1} \right) \boldsymbol{\mu}_{t\ell} - 2\boldsymbol{\mu}_{t\ell}^T \Sigma^{-1} \sum_i \boldsymbol{\eta}_i \right) \right] \times \exp\left[-\sum_i \frac{1}{2} \boldsymbol{\eta}_i^T \Sigma^{-1} \boldsymbol{\eta}_i \right] d\boldsymbol{\mu}_{tl}$$

$$= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} (2\pi)^{-pN_\ell/2} |\Sigma|^{-N_\ell/2} (2\pi)^{-p/2} |D|^{-1/2} \exp\left[-\sum_i \frac{1}{2} \boldsymbol{\eta}_i^T \Sigma^{-1} \boldsymbol{\eta}_i \right]$$

$$\times \int \exp\left[-\frac{1}{2} \left(\boldsymbol{\mu}_{t\ell} - (D^{-1} + N_\ell \Sigma^{-1})^{-1} \Sigma^{-1} \textstyle\sum_i \boldsymbol{\eta}_i \right)^T (D^{-1} + N_\ell \Sigma^{-1}) \left(\boldsymbol{\mu}_{t\ell} - (D^{-1} + N_\ell \Sigma^{-1})^{-1} \Sigma^{-1} \textstyle\sum_i \boldsymbol{\eta}_i \right) \right]$$

$$\times \exp\left[\frac{1}{2} (\Sigma^{-1} \textstyle\sum_i \boldsymbol{\eta}_i)^T (D^{-1} + N_\ell \Sigma^{-1}) \Sigma^{-1} \textstyle\sum_i \boldsymbol{\eta}_i \right] d\boldsymbol{\mu}_{tl}$$

$$\text{Kernal of MVN} \left((D^{-1} + N_\ell \Sigma^{-1})^{-1} \Sigma^{-1} \sum_i \boldsymbol{\eta}_i, (D^{-1} + N_\ell \Sigma^{-1})^{-1} \right)$$

$$
\begin{aligned}
&= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} (2\pi)^{-pN_\ell/2} |\Sigma|^{-N_\ell/2} (2\pi)^{-p/2} |D|^{-1/2} |2\pi (D^{-1} + N_\ell \Sigma^{-1})^{-1}|^{1/2} \\
&\qquad \times \exp\left[-\sum_i \frac{1}{2} \boldsymbol{\eta}_i^T \Sigma^{-1} \boldsymbol{\eta}_i \right] \exp\left[\frac{1}{2} (\Sigma^{-1} \sum_i \boldsymbol{\eta}_i)^T (D^{-1} + N_\ell \Sigma^{-1}) \Sigma^{-1} \sum_i \boldsymbol{\eta}_i \right] \\
&= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} (2\pi)^{-pN_\ell/2} |\Sigma|^{-N_\ell/2} |D|^{-1/2} |(D^{-1} + N_\ell \Sigma^{-1})^{-1}|^{1/2} \\
&\qquad \times \exp\left[-\frac{1}{2} \sum_i \boldsymbol{\eta}_i^T \Sigma^{-1} \boldsymbol{\eta}_i + \frac{1}{2} (\Sigma^{-1} \sum_i \boldsymbol{\eta}_i)^T (D^{-1} + N_\ell \Sigma^{-1}) \Sigma^{-1} \sum_i \boldsymbol{\eta}_i \right] \\
&= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} (2\pi)^{-pN_\ell/2} |\Sigma^{N_\ell} D (D^{-1} + N_\ell \Sigma^{-1})|^{-1/2} \\
&\qquad \times \exp\left[-\frac{1}{2} \sum_i \boldsymbol{\eta}_i^T \Sigma^{-1} \boldsymbol{\eta}_i + \frac{1}{2} (\Sigma^{-1} \sum_i \boldsymbol{\eta}_i)^T (D^{-1} + N_\ell \Sigma^{-1}) \Sigma^{-1} \sum_i \boldsymbol{\eta}_i \right]
\end{aligned}
$$

Thus, $\Lambda(\mathcal{T}_t)$ can be computed in closed form. Additionally, by conjugacy of the multivariate normal distribution, we have the full conditionals

$$
\boldsymbol{\mu}_{t\ell} \sim \mathcal{N}_k \left((D^{-1} + N_\ell \Sigma^{-1})^{-1} \Sigma^{-1} \sum_i \boldsymbol{\eta}_i,\ (D^{-1} + N_\ell \Sigma^{-1})^{-1} \right), \tag{A.7}
$$

where N_ℓ is the number of observations associated to leaf ℓ of tree t. We adopt the same Metropolis-Hastings steps for modifying the tree structure form as with `skewBART`.

A.6 MCMC for the `cBART` fitting

The conditional distribution of ψ_i upon ω_t, the contribution of observation i to the likelihood is

$$
\begin{aligned}
L(\psi_i) &\propto \exp\left\{ \kappa_i \psi_i - \omega_i \left(\psi_i^2 \right)^2 / 2 \right\} \\
&\propto \exp\left\{ -\frac{\omega_i}{2} \left(\frac{y_i - \xi}{2\omega_i} - \psi_i \right)^2 \right\}
\end{aligned}
$$

Given the values of $\{\omega_i\}_{i=1,N}$ and the prior, the conditional posterior of ψ can expressed as

$$(\psi \mid -) \propto \mathscr{N}\left(\boldsymbol{\psi}; \sum_{t=1}^{m} g(x, \mathcal{T}_t, \mathcal{M}_t), \tau\mathbf{I}\right) \prod_{i=1}^{N} e^{-\frac{\omega_i}{2}} \left(\psi_i - \frac{y_i - r}{2\omega_i}\right)^2$$

Then, Gibbs sampling proceeds in two simple steps:

$$(\omega_i \mid -) \sim PG\left(y_i + r, \psi_i\right)$$
$$(\boldsymbol{\psi} \mid -) \sim \mathscr{N}(\boldsymbol{\mu}_\psi, \boldsymbol{\Sigma}_\psi),$$

where $\Omega = \text{diag}\left(\omega_1 \cdots, \omega_N\right)$, $\Sigma_\psi = (\sigma^{-2}\mathbf{I} + \Omega)^{-1}$ and $\boldsymbol{\mu}_\psi = \Sigma_\psi \left((Y - r)/2 + \sigma^{-2} \sum_{t=1}^{T} g(X, \mathcal{T}_t, \mathcal{M}_t)\right)$

The r can be updated with a hybid Monte Carlo and a Metropolis-Hastings algorithms since the conditional posterior of r is proportional to $\prod_{i=1}^{N} \text{NB}(y_i; r, p_i)\, \text{Gamma}(r; a, b)$.

A.7 MCMC for the `mBART` fitting

Given the values of $\{\omega_i\}$, the conditional posterior of ψ can be expressed as

$$(\boldsymbol{\psi} \mid -) \propto \mathscr{N}\left(\mu_\psi, \Sigma_\psi\right) \prod_{i=1}^{N} e^{-\frac{\omega_i}{2}} \left(\psi_i - \frac{y_i - r}{2\omega_i}\right)^2$$

where $\mu_\psi = \sum_{t=1}^{T} G_3(x_i, \mathcal{T}_t, \mathcal{M}_t) + \Sigma_{13}\Sigma_{1:2}^{-1} \left(\boldsymbol{Y} - \sum_{t=1}^{T} G_{1:2}(x_i, \mathcal{T}_t, \mathcal{M}_t)\right)$ and $\Sigma_\psi = \Sigma_{33} - \Sigma_{13}\Sigma_{1:2}^{-1}\Sigma_{31}$ for $J = 3$ mixed responses.

Therefore,

$$(\boldsymbol{\psi} \mid -) \sim \mathscr{N}(\boldsymbol{\mu}^{post}, \boldsymbol{\Sigma}^{post}),$$
$$(\omega_i \mid -) \sim PG\left(y_i + r, \psi_i\right)$$

where $\boldsymbol{\mu}^{post} = \Sigma^{post}((y - r)/2 + 1/\Sigma_\psi \sum_{j=1}^{m} \boldsymbol{\mu}_\psi)$, $\Sigma^{post} = (1/\Sigma_\psi I + \Omega)^{-1}$ and $\Omega = \text{diag}\left(\omega_1 \cdots, \omega_N\right)$.

Update $\mathcal{T}$ and $\mathcal{M}$ where the dependent variable is $\boldsymbol{\psi}$.

Table A.1: Summary statistics of responses in GADD Study

Response	Description	Mean	SD	Min	Max
CAL	Clinical Attachment Level	2.06	1.13	0.31	9.33
PPD	Periodontal Pocket Depth	1.95	0.78	0.82	7.62
BOP	Bleeding on probing	58.78	37.09	0.00	155.00

A.8 Summary statistics of GADD study

Table A.2: Summary statistics of covariates in GADD Study

Variable	Description	Mean	SD	Min	Max
Age	patient age	55.27	10.67	26.00	87.00
Gender	patient gender (0=female, 1=male)	0.24	-	0	1
Bmi	overweight and obesity	35.33	9.01	6.08	66.25
Smoker	smoking status (0=Non-smoking, 1=smoking)	0.14	-	0	1
Hba1c	the amount of blood sugar attached to hemoglobin	7.92	2.03	5.00	16.40

BIBLIOGRAPHY

[1] R.B. Arellano-Valle, H. Bolfarine, and V.H. Lachos. Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 34(6):663–682, 2007.

[2] RB Arellano-Valle, H Bolfarine, and VH Lachos. Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 34(6):663–682, 2007.

[3] Reinaldo B. Arellano-Valle and Marc G. Genton. On fundamental skew distributions. *Journal of Multivariate Analysis*, 96(1):93–116, 2005.

[4] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

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

[6] Adelchi Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188, 2005.

[7] 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.

[8] Piyali Basak, Antonio R Linero, Debajyoti SInha, and Stuart Lipsitz. Semiparametric analysis of clustered interval-censored survival data using soft bayesian additive regression trees (sbart). *arXiv preprint arXiv:2005.02509*, 2020.

[9] 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.

[10] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[11] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[12] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regressiontrees. *The Annals of Applied Statistics*, 4(1):266 – 298, 2010.

[13] Siamak Zamani Dadaneh, Mingyuan Zhou, and Xiaoning Qian. Bayesian negative binomial regression for differential expression with confounding factors. *Bioinformatics*, 34(19):3349–3356, 04 2018.

[14] M. L. Darby and M. Walsh. Dental hygiene: Theory and practice. *Elsevier Health Sciences.*, 4 edn, 2014.

[15] 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.

[16] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal Japanese Society For Artificial Intelligence For Artificial Intelligence*, 14(771-780):1612, 1999.

[17] Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[18] Seymour Geisser and William F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

[19] Arjun K. Gupta, Truc T. Nguyen, and Jose Almer T. Sanqui. Characterization of the skew-normal distribution. *Annals of the Institute of Statistical Mathematics*, 56(2):351–360, 2004.

[20] Trevor Hastie and Robert Tibshirani. Bayesian backfitting. *Statistical Science*, 15(3):196–213, 2000.

[21] ME Herring and SK. Shah. Periodontal disease and control of diabetes mellitus. *The Journal of the American Osteopathic Association*, 106:416–421, 2006.

[22] Ida Johnson-Spruill, Pamela Hammond, Bertha Davis, Zina McGee, and Delroy Louden. Health of gullah families in south carolina with type 2 diabetes. *The Diabetes Educator*, 35(1):117–123, 2009.

[23] Adam Kapelner and Justin Bleich. bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software, Articles*, 70(4):1–40, 2016.

[24] Rico Krueger, Prateek Bansal, and Prasad Buddhavarapu. A new spatial count data model with bayesian additive regression trees for accident hot spot identification. *Accident Analysis & Prevention*, 144:105623, 2020.

[25] Xiaoyun Li, Dipankar Bandyopadhyay, Stuart Lipsitz, and Debajyoti Sinha. Likelihood methods for binary responses of present components in a cluster. *Biometrics*, 67(2):629–635, 2011.

[26] Yinpu Li, Antonio R Linero, and Jared S Murray. Adaptive conditional distribution estimation with bayesian decision tree ensembles. *arXiv preprint arXiv:2005.02490*, 2020.

[27] Antonio R. Linero. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018.

[28] Antonio R. Linero, Debajyoti Sinha, and Stuart R. Lipsitz. Semiparametric mixed-scale models using shared bayesian forests. *Biometrics*, 76(1):131–144, 2020.

[29] A.R. Linero and Y. Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society, Series B*, 80:1087–1110, 2018.

[30] Brian L. Mealey and Thomas W. Oates. Diabetes mellitus and periodontal diseases. *Journal of Periodontology*, 77(8):1289–1303, 2006.

[31] J. S. Murray. Log-linear bayesian additive regression trees for categorical and count responses. *arXiv preprint arXiv:*, 1701.01503, 2017.

[32] Jared S. Murray. Log-linear bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, 116(534):756–769, 2021.

[33] Jonathan Pillow and James Scott. Fully bayesian inference for neural models with negative-binomial spiking. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[34] Jonathan W. Pillow and James G. Scott. Fully bayesian inference for neural models with negative-binomial spiking. page 1898–1906, 2012.

[35] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

[36] 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.

[37] Veronika Ročková and Stéphanie van der Pas. Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108 – 2131, 2020.

[38] 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.

[39] Rodney A. Sparapani, Brent R. Logan, Robert E. McCulloch, and Purushottam W. Laud. Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, 35(16):2741–2753, 2016.

[40] Jennifer E. Starling, Jared S. Murray, Carlos M. Carvalho, Radek K. Bukowski, and James G. Scott. BART with targeted smoothing: An analysis of patient-specific stillbirth risk. *The Annals of Applied Statistics*, 14(1):28 – 50, 2020.

[41] Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 09 2012.

BIOGRAPHICAL SKETCH

Seungha Um was born in Yeongju, South Korea. She completed her Bachelor of Sciences degree in Applied Biosciences at Konkuk University in 2012. In 2016, she received her Master of Arts degree in Applied Statistics from Konkuk University. After receiving her master's degree, she pursued her Ph.D. in the Department of Statistics at the Florida State University. She defended her dissertation in the summer of 2021. Her primary research has focused on developing flexible Bayesian nonparametric models.