Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE

A monotone single index model for missing-at-random longitudinal proportion data

Satwik Acharyya[1] | Debdeep Pati[2] | Dipankar Bandyopadhyay[3] | Shumei Sun[3]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Statistics, Texas A&M University, College Station, TX 77840, USA
[3]Department of Biostatistics, Virginia Commonwealth University, Richmond VA 23298, USA

Correspondence
Dipankar Bandyopadhyay, Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, One Capitol Square, 7th Floor, PO Box 980032, Richmond, VA 23298-0032, USA. Tel: +1 804 827 2058; Fax: +1 804 828 8900 Email: dbandyop@vcu.edu

Summary

Beta distributions are commonly used to model proportion valued response variables, commonly encountered in longitudinal studies. In this article, we develop semi-parametric Beta regression models for proportion valued responses, where the aggregate covariate effect is summarized and flexibly modeled, using a interpretable monotone time-varying single index transform of a linear combination of the potential covariates. We utilize the potential of single index models, which are effective dimension reduction tools and accommodate link function misspecification in generalized linear mixed models. Our Bayesian methodology incorporates the missing-at-random feature of the proportion response, and utilize Hamiltonian Monte Carlo sampling to conduct inference. We explore finite-sample frequentist properties of our estimates, and assess the robustness via detailed simulation studies. Finally, we illustrate our methodology via application to a motivating longitudinal dataset on obesity research recording proportion body fat.

KEYWORDS:
Beta regression; Body fat; Proportion data; Single-index model; Monotone; Hamiltonian Monte Carlo

1 | INTRODUCTION

Research in various biomedical disciplines and public health generates data, where the primary response variables are constrained in a compact interval, say in $(0, 1)$, rather than the whole real line. For example, consider our motivating Fels longitudinal study (Roche 1992, FLS) recording proportion body fat (pbf), a popular clinical biomarker in obesity research, for study subjects at longitudinal time-points, along with various important covariates, such as gender, age, body mass index, etc. Figure 1a presents the raw density histogram of the response 'pbf' $\in (0, 1)$, packed across all subjects and time-points. For conducting regression, one can potentially transform the response variable to the real line, and use conventional approaches (Qiu, Song, & Tan 2008). However, such vanilla approaches pose both modeling and computational issues, given that inference can be sensitive to the transformations used, and parameters in the transformed scale rarely carry similar interpretation as in the original model. In a quest to conduct direct modeling of such responses, the Beta density (Gupta & Nadarajah 2004), a continuous log-concave density, is often the density of choice due to its versatility in accommodating a variety of unimodal shapes on a compact interval, and thereby address non-Gaussianity, and data skewness(Smithson & Verkuilen 2006). Under a generalized linear mixed model (GLMM) framework, a reparametrized beta density (and associated regression) assists us to conveniently connect the model covariates to the proportion response (Ferrari & Cribari-Neto 2004), with a subject-specific random effects in clustered/longitudinal studies (Hunger, Döring, & Holle 2012; Petterle, Bonat, & Scarpin 2019). Other popular distributions modeling proportion responses include the beta rectangular (Bayes, Bazán, & García 2012; Hahn 2008), simplex (Barndorff-Nielsen & Jørgensen 1991), logistic normal (Aitchison 1986), and the Bessel (Barreto-Souza, Mayrink, & Simas 2020).

Although linear regression (LR) are simple and commonly used procedures for evaluating covariate-response relationships, they are inadequate for inference and prediction under violations of the LR assumptions. In biomedical (obesity) research, index measures, such as the Charlson comorbidity index (Afolabi et al. 2020), combines information from an array of observed characteristics into a *single value*, thereby providing important unobserved traits of a subject (Wu & Tu 2016). However, a majority of these indices were developed on empirical grounds, and lacks sound statistical justification. On the other hand, a single-index model (Stoker 1986, SIM) provides a simple, interpretable framework for quantifying a complex, possibly non-linear relationship between a response Y_i and the $p > 1$ dimensional covariate vector $X_i = (X_{i1}, \ldots, X_{ip})$, where the conditional expectation of $Y_i|X_i$ can be expressed as an unknown, univariate function $g(.)$ of the scalar index $Z_i = X_i^T\beta$, where $\beta = (\beta_1, \ldots, \beta_p)$ is an unknown index vector (more details in Section 2). This SIM specification accommodates both non-linear main effects, and higher order interactions (determined by the function $g(.)$), and thus offer a pragmatic compromise between a fully parametric LR, and other nonparametric formulation (Dhara, Lipsitz, Pati, & Sinha 2020). However, (clinical) interpretation can be compromised(Foster, Taylor, & Nan 2013), if the shape of $g(.)$ is left completely unspecified. Hence, monotone single-index models (Balabdaoui, Durot, & Jankowski 2019; Groeneboom & Hendrickx 2019) have evolved, leading to straightforward clinical interpretation, and ease of inference. In this paper, we model the longitudinal proportion response using a BR, where the logit transform of its mean is flexibly modeled via a monotone single index transform of a linear combination of time-varying covariates. The justification behind enforcing monotonicity of the index function appears in the next section. Under a Bayesian paradigm powered by the Hamiltonian Monte Carlo (MMC) sampling (Betancourt, Byrne, Livingstone, & Girolami 2017), a missing-at-random (MAR) assumption accommodates seamless handling of the missing responses within the Bayesian updating scheme.

The rest of the article proceeds as follows: After a brief introduction to BR and the challenges associated with a SIM, Section 2 presents the monotone SIM specification via Bernstein polynomials (BP), and the consequent adaptation to handle MAR missingness. In Section 3, we develop the Bayesian estimation scheme, with prior specifications, posterior inference via HMC, and associated Bayesian model selection tools. Application to the motivating FELS data, with relevant posterior summaries and prediction accuracy appear in Section 4. In Section 5, we use synthetic data to evaluate the finite sample properties, and robustness of our proposal, in light of existing alternatives. Finally, some concluding comments and future developments appear in Section 6.

2 | STATISTICAL MODEL

2.1 | Single-index Beta Regression

Let y_{it} is the observed proportion response $\in (0, 1)$, and $X_{it} \in \mathbb{R}^p$ the corresponding predictors at tth $(t = 1, \ldots, T_i)$ time point for the ith $(i = 1, \ldots, n)$ subject. We model y_{it}, conditional on the covariates as

$$f_{BR}(y_{it} \mid \mu_{it}, \psi_i) = \frac{\Gamma(\psi_i)}{\Gamma(\mu_{it}\psi_i)\Gamma((1-\mu_{it})\psi_i)} y_{it}^{\mu_{it}\psi_i - 1}(1-y_{it})^{(1-\mu_{it})\psi_i - 1}, \quad 0 < \mu_{it} < 1, \psi_i > 0. \tag{1}$$

where, $\mu_{it} = E(y_{it})$ is the mean parameter, with individual specific precision parameter ψ_i. This is denoted as $y_{it} \sim \text{Beta}(\mu_{it}\psi_i, (1-\mu_{it})\psi_i)$. Now, to propose the beta regression, the covariates are connected to the mean μ_{it} as $\text{logit}(\mu_{it}) = X_{it}^T\beta$, where $\beta \in \mathbb{R}^p$ is the vector of regression coefficients. The variance component ψ_{it} is left unspecified (to be estimated using some priors as in our case), or estimated via assigning a link function, such as log, to the covariates.

Over the years, there has been considerable effort in constructing more flexible mean functions of multivariate covariates. In this context, the SIMs(Hardle, Hall, & Ichimura 1993), where $\text{logit}(\mu_{it})$ is expressed as $g(X_{it}^T\beta)$, is typically viewed as a bridge between a multiple linear regression, and a non-parametric regression problem. The SIM do not suffer from the curse of dimensionality due to the reduction to an univariate index variable from the multidimensional predictor set (Yu & Ruppert 2002). This dimension reduction property enhances computational scalability, and preserves the flexibility of the model through utilization of non-linear functions. They are also advantageous in the context of misspecification of the non-linear link function on $X_{it}^T\beta$. The function $g(\cdot)$ is called a link function, and the p-variate coefficient vector β is called the index vector. The linear combination $X_{it}^T\beta$ is referred as the index for the predictor X_{it}. In a SIM, one aims to estimate both the link function g and the index vector β simultaneously. The problem of estimating g is the same as a non-parametric univariate regression problem. An advantage of using a SIM over a usual multiple linear regression problem is that once can achieve a higher predictive power as we are considering a function of a linear combination of the covariates (Carroll, Fan, Gijbels, & Wand 1997). SIM is also a special case of projection pursuit regression (PPR) (Friedman & Stuetzle 1981) with a single component, thereby providing simpler interpretation over the multi-component PPR. Thus, in addition to better prediction performance, the SIM provides appealing interpretation of the covariate effects on the response.

Despite its popularity, the SIM brings in certain challenges for statistical inference. The parameters (g, β) are not jointly identified. Constraints on (g, β), such as monotonicity on g (Foster et al. 2013), and a unit norm constraint on β facilitate identifiability and the interpretation of the covariate effect on the response. With $g(.)$ assumed monotone, there are additional advantages in interpretation and usefulness of the index $Z_{it} = X_{it}^T\beta$ for subject i. Without loss of generality, if g is monotone non-decreasing, the expected value of the response will increase or remain the same with

increase in Z_{it}, and a prespecified threshold on $g(X_i^T\beta)$ enjoys a one-on-one correspondence to an equivalent threshold on $X_i^T\beta$, allowing elegant interpretation. Moreover, if the coefficient/parameter corresponding to a specific X_{it} is positive, then increasing the value of X_{it} (with other covariates remaining fixed) will result in a higher value of the index, thus increasing the expected value of the response. Based on these interpretations, clinicians may aim to lower the expected adverse response by taking appropriate steps to lower the index, Z_{it}, by implementing a plan to decrease/increase a particular X_{it}. These physical interpretations of g and β are not always available when the link function g is unconstrained. In the following, we present the details of our monotone SIM for longitudinal proportion data, and the adjustments to handle missingness.

2.2 | Monotone SIM

Classical methods (Ayer, Brunk, Ewing, Reid, & Silverman 1955; Brunk 1955) of imposing monotonicity requires constrained optimization to ensure the monotonicity and smoothness of the link function. Isotonic regression i.e, fitting a monotone function g to data points (y_i, x_i) involves finding a weighted least-squares fit $p \in \mathbb{R}^n$ to the vector $y \in \mathbb{R}^n$, with weights vector $w \in \mathbb{R}^n$ subject to a set of constraints of the kind $p_i \leq p_{i+1}$. Then, the isotonic regression problem corresponds to the following quadratic program: $\min \sum_{i=1}^n w_i(p_i - y_i)^2$ subject to $p_i \leq p_{i+1}$. This can be solved using a simple iterative algorithm called the pool adjacent violators algorithm (De Leeuw, Hornik, & Mair 2010). In presence of the single index, such optimization routines can be problematic.

To circumvent this issue, we use a smooth BP basis (Farouki 2012) to model the link function. In the following, we describe how using a BP basis reduces the problem of estimating a monotone link function to a constrained linear regression problem. In a Bayesian context, optimization is avoided by placing suitable prior distributions on the constrained space, and then sampling from the posterior distribution to produce estimates that satisfy the constraints. The SIM connects the linear predictors with a non-linear function to model the logit linked mean of the response variable as

$$\text{logit}(\mu_{it} \mid z_i) = g(X_{it}^T\beta) + z_i, \quad z_i \sim N(0, \sigma_z^2). \tag{2}$$

where $g(.)$ is a unknown monotone link function on $\mathbb{R} \to \mathbb{R}$, and z_i denote subject-specific random effects. We impose a standard restriction on β i.e. $||\ \beta\ || = 1$, to ensure identifiability of the model. Detailed discussion appears in Subsection 2.3. This assumption helps to provide a better interpretation of the response variable i.e., pbf, wrt. the time component. We model the monotone link function $g(\cdot)$ with BP and the monotonicity is ensured through imposing necessary restriction on the coefficients of the polynomial. The BP of degree M is defined as

$$B_M(v) = \sum_{j=0}^{M} \theta_j B_{M,j}(v), \quad B_{M,j}(v) = \binom{M}{j} v^j (1-v)^{M-j}, \quad j = 0, \ldots, M,\ v \in [0,1]. \tag{3}$$

where, $B_M(v)$ is non-decreasing if the coefficients of the polynomial are non-decreasing i.e. $\theta_0 < \theta_2 < \cdots < \theta_M$ (Chak, Madras, & Smith 2005). Following Souris, Bhattacharya, and Pati (2018), we scale the input variable $X_{it}^T\beta$ because the domain of BP is defined on $[0,1]$ interval. Using Cauchy-Schwarz inequality and identifiability constraints, we have $|\ X_{it}^T\beta\ | \leq ||\ X_{it}\ ||\ ||\ \beta\ || = ||\ X_{it}\ ||$ where $||\ \beta\ || = 1$. We consider $c = \max ||\ X_{it}\ ||$, and transform $\tilde{X}_{it} = X_{it}/c_i$, leading to the identity $|\ \tilde{X}_{it}^T\alpha\ | \leq 1$. Another transformation is required on the BP $B_{M,j}(v) = p_j(u)/(M+1), v \in [0,1]$ to provide support on $[-1,1]$, where, $p_j(v)$ is a Beta$(j+1, M-j+1)$ density. We take the transformation $W = 2V - 1$, where $p_j(v)$ is the density of the variable V and the density of W is $q_j(w) = p_j\{(w+1)/2\}/2$ for $w \in [-1,1]$. The transformed BP basis for $j = 0, \ldots, M$ is then defined as

$$\tilde{B}_{M,j}(w) = q_j(w)/(M+1) \qquad w \in [-1,1].$$

The monotone SIM with transformed BP basis is given as

$$g(\tilde{X}_{it}^T\beta) = \tilde{B}_M(\tilde{X}_{it}^T\beta) = \sum_{j=0}^{M} \theta_j \tilde{B}_{M,j}(\tilde{X}_{it}^T\beta), \qquad |\ \tilde{X}_{it}^T\alpha\ | \leq 1 \tag{4}$$

where $\tilde{X}_{it} = X_{it}/\max ||\ X_{it}\ ||$, $\tilde{B}_{M,j}(w) = q_j(w)/(M+1)$ and $q_j(w)$ is a transformed beta density with $|\ w\ | < 1$. Now, $\tilde{B}_M(.)$ is non-decreasing because of the order-restriction on the basis coefficients $\{\theta_j\}_{j=0}^M$. We define an equivalent transformation on $\{\theta_j\}_{j=1}^M$ and set $\phi_0 = \theta_0, \phi_1 = \theta_1 - \theta_0, \ldots, \phi_M = \theta_M - \theta_{M-1}$, such that $\phi_k \geq 0$ for $k = 1, \ldots, M$. We write $A\Phi = \theta$ where $\theta = [\theta_0, \theta_1, \ldots, \theta_M]$, $\Phi = [\phi_0, \phi_1, \ldots, \phi_M]$ and A is a $(M+1) \times (M+1)$ dimensional matrix with all the lower triangle and diagonal entries are 1. Thus, (2) can be rewritten as

$$\text{logit}(\mu) = \mathbb{B}_\alpha\theta + z = \mathbb{B}_\alpha A\Phi + z \tag{5}$$

where $\mathbb{B}_\alpha = [\tilde{B}_M^{11}, \ldots, \tilde{B}_M^{nT_n}]^T$ is a $nT \times (M+1)$ matrix with $T = \sum_{i=1}^n T_i$ and $\tilde{B}_M^{it} = [\tilde{B}_{M,0}(\tilde{X}_{it}^T\beta), \ldots, \tilde{B}_{M,M}(\tilde{X}_{it}^T\beta)]^T$.

2.3 | Identifiability of parameters

Lin and Kulasekera (2007) proved the identifiability constraints for the SIM, under the assumption of non-constant and continuous g, $\|\beta\| = 1$, with the first non-zero element being positive. In our case, the monotonicity property of the function g assists in relaxing the assumption of the first

non-zero element of β to be positive (Balabdaoui et al. 2019). The proof of the identifiability constraint is also extended to left or right continuous functions instead of continuous functions (Balabdaoui et al. 2019) under i.i.d. setup. Observe that due to the presence of random effects, the well-studied iid setup has been violated. However, this does not pose a significant problem to show identifiability as we shall see below.

To define the model (2), we need at least one observation from a subject at one time point. We define the support of g as $\mathcal{C}_\beta = \{x^T\beta, x \in \mathbb{R}^p\}$. We assume that mean of the response variable exists and (2) holds true for some $\gamma = (g, \beta)$ such that $\beta \in \mathcal{S}_{p-1}$ (a unit sphere of dimension p) and $g \in \mathcal{M}$, the class of monotone functions on $\mathbb{R}$. We enforce identifiability of the parameters through imposing the following constraints: (a) $g(.)$ is a monotone non-decreasing function i.e., $g \in \mathcal{M}$, and (b) $|| \beta ||= 1$ i.e. $\beta \in \mathcal{S}_{p-1}$. To demonstrate identifiability, we assume that there exists parameters $\gamma_1 = (g_1, \beta_1)$ and $\gamma_2 = (g_2, \beta_2)$, $g_i \in \mathcal{M}_{p-1}$, $\beta_i \in \mathcal{S}_{p-1}$, such that $f(y_{11}, x_{11} \mid \gamma_1) = f(y_{11}, x_{11} \mid \gamma_2)$. Taking expectation of (2) and applying the inverse logit transformation, we have $g_1(x^T\beta_1) = g_2(x^T\beta_2)$. To uniquely identify the parameters, it is sufficient to show $g_1(x^T\beta_1) = g_2(x^T\beta_2)$, which only holds if $\beta_1 = \beta_2, g_1 \equiv g_2$. Our final claim holds true by Proposition 5.1 from Balabdaoui et al. (2019).

2.4 | Handling Missingness

Missing data or incomplete information is a common issue in medical studies. Several techniques to handle missing data have been studied over last few decades using approaches such as data imputation (Harel & Zhou 2007; Rubin 2004; Zhang 2003) and fully Bayes (Daniels & Hogan 2008; Ibrahim, Chen, Lipsitz, & Herring 2005). In absence of a proper clinical justification for missing-not-at-random assumption, we assume the response pbf to be missing at random (MAR), also referred to as ignorable missingness (Little & Rubin 2019). This implies that the missing data mechanism is not dependent on the missing response values.

It is an established fact that unbiased estimates can be obtained from the observed likelihood instead of the joint likelihood of observed and missing data (Seaman, Galati, Jackson, & Carlin 2013). We denote R_{it} as the indicator for the missing response variables at tth time point for the ith individual i.e. $R_{it} = 1$ if Y_{it} is observed, 0 otherwise. The conditional distribution of missing data mechanism $f(R \mid Y, \lambda)$ is identified through the parameter vector λ which is independent from our parameter of interest Θ. In case of MAR, the conditional distribution is independent of the choice of missing response values. Following the definition from Rubin (1976), MAR holds if

$$f(R = r \mid Y_{obs}, Y_{mis}, \lambda, \Theta) = f(R = r \mid Y_{obs}, \lambda, \Theta)$$

where Y_{obs} and Y_{mis} are observed and missing set of responses, along with λ, the parameter vector of the missingness mechanism. The conditional distribution of $(Y_{obs}, X, R \mid \lambda, \Theta)$ is

$$\begin{aligned}
f(Y_{obs}, X, R \mid \lambda, \Theta) &= \int f(Y_{obs}, Y_{mis}, X, R|\lambda, \Theta) dY_{mis} \\
&= \prod_{i,t|R_{it}=1} f(Y_{obs,it}, X_{it}, R_{it} = 1|\lambda, \Theta) \int \prod_{i,t|R_{it}=0} f(Y_{mis,it}, X_{it}, R_{it} = 0|\lambda, \Theta) dY_{mis} \\
&= \prod_{i,t|R_{it}=1} f(Y_{obs,it}, X_{it}, R_{it} = 1|\lambda, \Theta) \prod_{i,t|R_{it}=0} f(X_{it}) \int f(R_{it} = 0|X_{it}, \lambda) f(Y_{mis,it} \mid X_{it}, \Theta) dY_{mis} \\
&= \prod_{i,t|R_{it}=1} f(Y_{obs,it}, X_{it}, R_{it} = 1|\lambda, \Theta) \prod_{i,t|R_{it}=0} f(X_{it}) f(R_{it} = 0|X_{it}, \lambda).
\end{aligned} \tag{6}$$

Under MAR assumptions, equation (6) holds, clearly indicating that the inference on Θ (parameter of interest) does not depend on the missing data mechanism. A Bayesian approach can naturally incorporate the uncertainty due to the presence of missing data (Erler et al. 2016). Bayesian methods on missing data can generate posterior samples of the parameters and missing variables from their posterior predictive distribution. In the next section, we outline our hierarchical Bayesian estimation framework that incorporates the missing data model.

3 | BAYESIAN INFERENCE

3.1 | Prior specification

Our Bayesian estimation scheme is initiated via specifying the prior distributions on the model parameters. First, we denote $\alpha = \beta / || \beta ||$ and place a Gaussian prior on β. Next, we posit a standard Gaussian prior on the first coordinate of Φ, while the remaining coordinates together gets a multivariate Gaussian prior $N(0, cI_M)$ truncated to a positive real line. A non-informative prior is imposed on the variance of the subject-specific random effects, i.e. σ_z^2 in (2). Following the recommendations from Bandyopadhyay, Galvis, and Lachos (2017), we use a Gamma prior on the precision parameter ψ for the BR model in (1).

3.2 | Posterior Inference

Under the assumptions of independence between the subject-specific random effects and MAR, the observed likelihood of $\Theta = (\beta, \Phi, \psi, \sigma_z)$ is

$$\begin{aligned} L(\Theta, Z \mid y, X) &= \prod_i^n \left[\prod_{t=1}^{T_i} f_{BR}(y_{it} \mid \mu_{it}, \psi_i) \right]^{R_{it}=1} p(z_i) \\ &= \prod_i^n \left[\prod_{t=1}^{T_i} \frac{\Gamma(\mu_{it})}{\Gamma(\mu_{it}\psi_i)\Gamma((1-\mu_{it})\psi_i)} y_{it}^{\mu_{it}\psi_i - 1} (1-y_{it})^{(1-\mu_{it})\psi_i - 1} \right]^{R_{it}=1} p(z_i), \end{aligned} \tag{7}$$

where $p(z_i)$ denotes the distribution of subject-specific random effects, and μ_{it} is presented in (2). The joint posterior distribution can be written as

$$\pi(\Theta, Z \mid y, X) \propto L(\Theta, Z \mid y, X) \times \pi(\beta, \Phi, \psi, \sigma_z). \tag{8}$$

A standard technique to obtain posterior samples is via the implementation of the Markov chain Monte Carlo (MCMC) algorithm, which cycles through the full conditionals of a parameter given the rest. Instead of using standard MCMC algorithm, we obtain posterior samples by using Hamiltonian Monte Carlo, or HMC (Duane, Kennedy, Pendleton, & Roweth 1987; Neal 1994) and the No-U-turn sampler (NUTS) (Hoffman & Gelman 2014). A probabilistic programming language `Stan` (Carpenter et al. 2017; Stan Development Team 2019) has been developed for Bayesian inference by combining the HMC and NUTS sampler. `Stan` is scalable for large datasets, and often achieves faster convergence compared to other available software, such as `WinBUGS` (Lunn, Spiegelhalter, Thomas, & Best 2009), `JAGS` (Plummer 2003), and others. We fit our BR model (1) with `Stan` by specifying the observed likelihood (7) and prior distributions (3.1).

3.3 | Bayesian model selection and influence diagnostics

To assess model goodness of fit, we compared prediction accuracy through several diagnostic measures (Hoeting, Madigan, Raftery, & Volinsky 1999; Vehtari & Ojanen 2012). Such measures can be either information based, or cross-validatory. The Bayesian information criteria (Schwarz 1978, BIC) penalizes the log-likelihood with number of fitted parameters and sample size. WAIC (Gelman, Hwang, & Vehtari 2014) is also another popular fully Bayes information criteria. Computationally efficient WAIC estimates are obtained from out-of-sample prediction with pointwise log posterior predictive density along with an adjustment of number of effective parameters. Following Gelman et al. (2014), the effective number of parameters is computed as

$$p_{WAIC} = 2 \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left(\log \left(\frac{1}{S} \sum_{s=1}^{S} \pi(y_{it} \mid \Theta^s) \right) - \frac{1}{S} \sum_{s=1}^{S} \log \pi(y_{it} \mid \Theta^s) \right)$$

Finally, WAIC is evaluated as

$$WAIC = -2(lppd - p_{WAIC}) \tag{9}$$

where, lppd is log of pointwise predictive density.

Cross-validation (CV) based criteria are another parallel strategies to capture out-of-sample prediction errors. The diagnostic measures based on CV do not suffer from the problem of over-fitting but they may not be computationally scalable. The conditional predictive ordinate (Dey, Chen, & Chang 1997, CPO) criterion is calculated for an observed point given all other data points. The CPO for the i-th subject at the t-th time point is defined as $CPO_{it} = \pi(y_{it} \mid y_{(it)}) = \int \pi(y_{it} \mid \Theta)\pi(\Theta \mid y_{(it)})d\Theta$, where $y_{(it)}$ denotes the dataset without the observation y_{it}. Following Dey et al. (1997), CPO_{it} is numerically computed as

$$CPO_{it} = \left(\frac{1}{S} \sum_{s=1}^{S} \frac{1}{\pi(y_{it} \mid y_{(it)}, \Theta^s)} \right)^{-1} \tag{10}$$

where, S denotes the number of posterior Monte Carlo samples, post-convergence. A higher value of CPO_{it} for a model indicates a better support for the (i, t) th datum. A summary measure is the log pseudo-marginal likelihood (Geisser & Eddy 1979, LPML), defined as

$$LPML = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \log(CPO_{it}) \tag{11}$$

Similar to CPO, a higher value of LPML suggests a better model fit to the data. Under the MAR assumption, we evaluated the model diagnostic measures based only on observed responses (Daniels & Hogan 2008; Ma & Chen 2018).

4 | APPLICATION: FLS DATA

In this section, we illustrate our BR monotone SIM (BR-MSIM) via application to the motivating FLS data. The FLS (Sun et al. 2007 2008) is the world's longest and largest longitudinal human growth study that collected the lifetime of repeated measurements on growth, health and body

composition of 2,567 European-American participants, as early as 1929. Although participants were enrolled at birth (examined semi-annually until 18 years of age, and biennially thereafter), our current analytical data subset consists of 777 subjects (373 male and 404 females), followed longitudinally since 1976 (year when pbf measurements were included in the study protocol) till 2010. Study subjects exhibit irregular number of time points, with a maximum number of 15 visits. In addition to the response variable (pbf) $\in (0, 1)$ collected at each time-point, various subject-level covariates, such as gender (Gender, M/F), date of visits (Visit), age (Age, range = 8–83), body mass index (BMI), waist size (Waist), diastolic blood pressure (Dias BP), systolic blood pressure (Sys BP), bicep size(Bicep, in mm), and bio-electrical impedance (BCimped), were also available. Approximately, 17.8% of observations were missing, which we considered MAR. The study was approved by the Institutional Review Boards of the Wright State University and the Virginia Commonwealth University.

We compared the fit of the BR-MSIM to the BR model with linear predictors (BR-Lin), where both models vary with respect to having a subject-specific precision parameter (ψ_i), or an overall precision parameter (ψ). The competing models are listed below:

Model 1: $y_{it} \sim$ BR-MSIM(μ_{it}, ψ_i),

Model 2: $y_{it} \sim$ BR-Lin(μ_{it}, ψ_i),

Model 3: $y_{it} \sim$ BR-Lin(μ_{it}, ψ),

Model 4: $y_{it} \sim$ BR-MSIM(μ_{it}, ψ).

We choose the optimal value of $M \in \{5, \ldots, 30\}$ for models 1 and 4 based on BIC values. More details about the selection M is deferred to the section S2 of the Supplementary Material. Using the optimal value of $M = 22$, we compare the four models via WAIC and LPML values (Table 1), calculated using observed data (Ma & Chen 2018). We observe that Model 4 (BR-MSIM, with constant precision parameter ψ) has the highest LPML and the lowest WAIC values among the 4 competing models. The BR-MSIM with subject-specific precision ψ_i is performing poorly because of over-parametrization.

Next we report findings from our best-fitted Model 4, i.e., BR-MSIM(μ_{it}, ψ). The estimated posterior mean of regression coefficients with corresponding 95% credible intervals (CI) are presented in Figure 2a. The covariates BMI, Waist, Bicep, and BCimped exhibit significantly higher estimates, compared to the others. All covariates (except Age and Sys BP) have positive regression coefficients. This implies an increase in the estimated single index with an increase in the numerical value for the continuous covariates, or change in category (say, from 0 to 1) for the discrete covariates. However, the covariates Age and Sys BP negatively impact the single index.

Figure 2b plots the estimated single index, with the corresponding 95% CI, denoted by the grey area. To illustrate the monotonicity property of the nonparametric function $g()$, consider the single index w_{it} for the i-th subject at the t-th time point given by $w_{it} = X_{it}^T \beta$, and $s_{it} = g(w_{it})$. Due to this property, we have $s_{it_1} > s_{it_2}$ when $w_{it_1} > w_{it_2}$, where t_1 and t_2 are time points of two arbitrary visits of the ith subjects. For an illustration using the FLS data, consider the subject with $id = 8$ in the FLS data, who is a male with $t = 1, \ldots, 7$. For this subject, the values of the single index w at the 6th and 7th time-points are $w_{86} = 0.18$ and $w_{87} = 0.30$, with the corresponding $s_{86} = 0.66$ and $s_{87} = 0.85$, respectively.

Other than WAIC and LPML values, we also used exploratory graphs to assess the goodness of fit. We generate multiple samples from the predictive distribution using the posterior samples. We average over multiple samples to obtain the predicted values of the response variables. Figure 1c represents the association between the predicted values of response variable on the y-axis and observed response variable on the x-axis. The linear trend in the Figure 1c suggests an adequate for the BR-MSIM(μ_{it}, ψ) model. Next, we calculate prediction accuracy metrics, such as sum of squared errors (SSE), sum of absolute error (SAE), mean absolute percentage error (MAPE), and mean arctangent absolute percentage error (Kim & Kim 2016, MAAPE) to compare the four competing models. These prediction accuracy measures are summarized in Table 1. We observe that Model 4, i.e., the BR-MSIM(μ_{it}, ψ), has the lowest values corresponding to all metrics, implying superior prediction accuracy compared to the rest three models.

5 | SIMULATION STUDIES

In this section, we use synthetic data to (a) assess the frequentist finite sample properties (Simulation 1), and (b) assess robustness (Simulation 2), of our proposed monotone SIM.

5.1 | Simulation 1: Checking frequentist finite sample properties

Here, we investigate the consistency of single index parameters for increasing values of $n \in \{100, 200, 300, 400, 500\}$, while setting the percentage of missingness at 20%. To mimic a realistic setting as observed in the FLS electronic health records, varying number of observed (longitudinal) time points are generated via random sampling (with replacement) from $\{1, \ldots, 10\}$, with the maximum set to 10. We consider the dimension of regression and basis coefficient to be $p = 4$ and $M = 22$ respectively, and fix the true values of the regression coefficients $\beta = \{2.75, 0.85, 2, 1.25\}$,

normalized to have unit norm. We set $\psi = 3$, and the basis coefficients ϕ are sampled from a vector $(0, 0.05, 0.1, 0.2, 0.3, 0.4)$, with first entry fixed at 0.15. Posterior estimates were summarized over 50 replicates.

In Figure 3a, we plot the estimated single index function i.e. $g(\text{single index})$ and the associated 95% CIs for $n = 100$, with the truth overlayed. Under the same setting, we provide the plot for $n = 500$ in Figure 3b. Furthermore, we split the interval $[0, 1]$ with an increment of 0.01, and measured the quantiles of observed and estimated mean (μ_{it}) at those probabilities. Figures 3c and 3d plot the observed vs estimated quantiles from the mean. To check the consistency of the single index function $[g(\text{single index})]$ parameters, we define a measure of discrepancy as $d_s(a, b) = \frac{1}{N} \mid\mid a - b \mid\mid_2$, where $a, b \in \mathbb{R}^N$. Figure S3a (Supplementary material) presents the boxplots of $100 \times d_s$ between true and estimated single index function for increasing values of $n \in \{100, 200, 300, 400, 500\}$ across the 50 replicates. In Figure S3a, the decreasing trend (of Euclidean distances) in the boxplots with increasing n implies consistency of the posterior estimates of single index function parameters. We also report the average bias, MSE and 95% CIs of the regression coefficients β in Table 2. All three metrics decreases with increase in the sample size. To ensure convergence of the posterior samples, we also provide several trace plots in figure S1 of the Supplementary Material.

5.2 | Simulation 2: Assessing robustness

To assess the robustness of our model, we generate the response variable from a mixture of beta and simplex distribution (Barndorff-Nielsen & Jørgensen 1991), given as $(0.8 \times \text{Beta} + 0.2 \times \text{simplex})$, and fit our proposed BR-MSIM. We present similar plots for this misspecified case, as in Subsection 5.1. Figures 4a and 4b plot the true and estimated single index (with 95% CIs) for $n = 100$ and $n = 500$ respectively, when the data is generated from the misspecified data generating distribution. Contrary to Figure S3a, here, we observe an increasing trend in the boxplots of the Euclidean distances ($100 \times d_s$) with increasing n; see, Figure S3b in the Supplementary Material. The quantile plot of true and estimated mean of the response variable is provided in Figure 4c (n=100) and 4d ($n = 500$). The effect of misspecification is clear, with the quantiles moving away from the $y = x$ line.

6 | CONCLUSIONS

In this paper, we provide a unique methodology for monotone single index modeling under BR using Bernstein polynomials. This methodology can be extended for any distribution which is supported on the interval $(0, 1)$, such as beta rectangular (Bayes et al. 2012; Hahn 2008), simplex (Barndorff-Nielsen & Jørgensen 1991), logistic normal (Aitchison 1986). Our current setup assumes MAR missingness; certainly, this can be extended to MNAR missingness via the popular shared random effects framework (Albert & Follmann 2003) that jointly models the response variable and the (binary) missingness indicator.

ACKNOWLEDGMENTS

The research was supported by grants R01AG048801, R01DE024984, and P30CA016059 awarded by the United States National Institutes of Health.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:

References

Afolabi, H. A., bin Zakariya, Z., Shokri, A. B. A., Hasim, M. N. B. M., Vinayak, R., Afolabi-Owolabi, O. T., & Elesho, R. F. (2020). The relationship between obesity and other medical comorbidities. *Obesity Medicine*, *17*, 100164.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, Ltd.

Albert, P. S., & Follmann, D. A. (2003). A random effects transition model for longitudinal binary data with informative missingness. *Statistica Neerlandica*, *57*(1), 100–111.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, 641–647.

Balabdaoui, F., Durot, C., & Jankowski, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli*, *25*(4B), 3276–3310.

Bandyopadhyay, D., Galvis, D. M., & Lachos, V. H. (2017). Augmented mixed models for clustered proportion data. *Statistical methods in medical research*, *26*(2), 880–897.

Barndorff-Nielsen, O. E., & Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of multivariate analysis*, *39*(1), 106–116.

Barreto-Souza, W., Mayrink, V. D., & Simas, A. B. (2020). *Bessel regression model: Robustness to analyze bounded data*.

Bayes, C. L., Bazán, J. L., & García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, *7*(4), 841–866.

Betancourt, M., Byrne, S., Livingstone, S., & Girolami, M. (2017). The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, *23*(4A), 2257–2298.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 607–616.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017, 1). Stan : A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi: 10.18637/jss.v076.i01

Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.

Chak, P. M., Madras, N., & Smith, B. (2005). Semi-nonparametric estimation with Bernstein polynomials. *Economics Letters*, *89*(2), 153–156.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press.

De Leeuw, J., Hornik, K., & Mair, P. (2010). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, *32*(5), 1–24.

Dey, D. K., Chen, M.-H., & Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, 1239–1252.

Dhara, K., Lipsitz, S., Pati, D., & Sinha, D. (2020). A new Bayesian single index model with or without covariates missing at random. *Bayesian Analysis*, *15*(3), 759–780.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216–222.

Erler, N. S., Rizopoulos, D., Rosmalen, J. v., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in medicine*, *35*(17), 2955–2974.

Farouki, R. T. (2012). The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, *29*(6), 379–419.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Foster, J. C., Taylor, J. M., & Nan, B. (2013). Variable selection in monotone single-index models via the adaptive LASSO. *Statistics in Medicine*, *32*(22), 3944–3954.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*(376), 817–823.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997–1016.

Groeneboom, P., & Hendrickx, K. (2019). Estimation in monotone single-index models. *Statistica Neerlandica*, *73*(1), 78–99.

Gupta, A. K., & Nadarajah, S. (2004). *Handbook of Beta Distribution and its Applications*. CRC press.

Hahn, E. D. (2008). Mixture densities for project management activity times: A robust approach to pert. *European Journal of operational research*, *188*(2), 450–459.

Hardle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 157–178.

Harel, O., & Zhou, X.-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, *26*(16), 3057–3077.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382–401.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hunger, M., Döring, A., & Holle, R. (2012). Longitudinal beta regression models for analyzing health-related quality of life scores over time. *BMC medical research methodology*, *12*(1), 144.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*(469), 332–346.

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679.

Lin, W., & Kulasekera, K. (2007). Identifiability of single-index models and additive-index models. *Biometrika*, *94*(2), 496–501.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The bugs project: Evolution, critique and future directions. *Statistics in medicine*, *28*(25), 3049–3067.

Ma, Z., & Chen, G. (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, *47*(3), 297–313.

Neal, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, *111*(1), 194–203.

Petterle, R. R., Bonat, W. H., & Scarpin, C. T. (2019). Quasi-beta longitudinal regression model applied to water quality index data. *Journal of Agricultural, Biological and Environmental Statistics*, *24*(2), 346–368.

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, pp. 1–10).

Qiu, Z., Song, P. X.-K., & Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, *35*(4), 577–596.

Roche, A. F. (1992). *Growth, maturation, and body composition: the fels longitudinal study 1929-1991* (No. 9). Cambridge University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by" missing at random"? *Statistical Science*, 257–268.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54.

Souris, A., Bhattacharya, A., & Pati, D. (2018). The soft multivariate truncated normal distribution. *arXiv preprint arXiv:1807.09155*.

Stan Development Team. (2019). *RStan: the R interface to Stan.* Retrieved from `http://mc-stan.org/` R package version 2.19.1.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica: Journal of the Econometric Society*, 1461–1481.

Sun, S. S., Grave, G. D., Siervogel, R. M., Pickoff, A. A., Arslanian, S. S., & Daniels, S. R. (2007). Systolic blood pressure in childhood predicts hypertension and metabolic syndrome later in life. *Pediatrics*, *119*(2), 237–246.

Sun, S. S., Liang, R., Huang, T. T.-K., Daniels, S. R., Arslanian, S., Liu, K., … Siervogel, R. M. (2008). Childhood obesity predicts adult metabolic syndrome: the Fels Longitudinal Study. *The Journal of pediatrics*, *152*(2), 191–200.

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.

Wu, J., & Tu, W. (2016). A multivariate single-index model for longitudinal data. *Statistical Modelling*, *16*(5), 392–408.

Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*(460), 1042–1054.

Zhang, P. (2003). Multiple imputation: theory and method. *International Statistical Review*, *71*(3), 581–592.

FIGURES

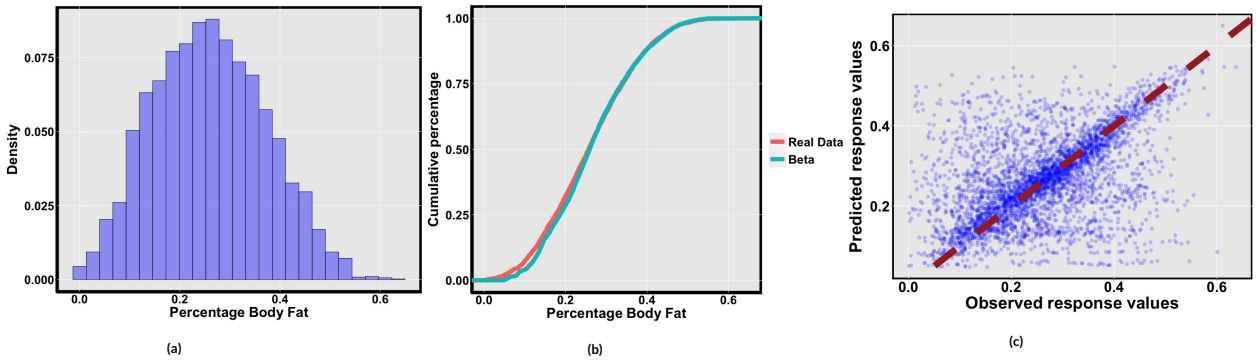

FIGURE 1 *The left panel shows the histogram of percentage body fat (pbf). The middle panel provides empirically calculated cumulative distribution of the response variable i.e. percentage body fat and the same after fitting beta regression combined with monotone single index* (2)*. The right panel plots the predicted vs observed response variables from BR-MSIM*(μ_{it}, ψ) *model.*

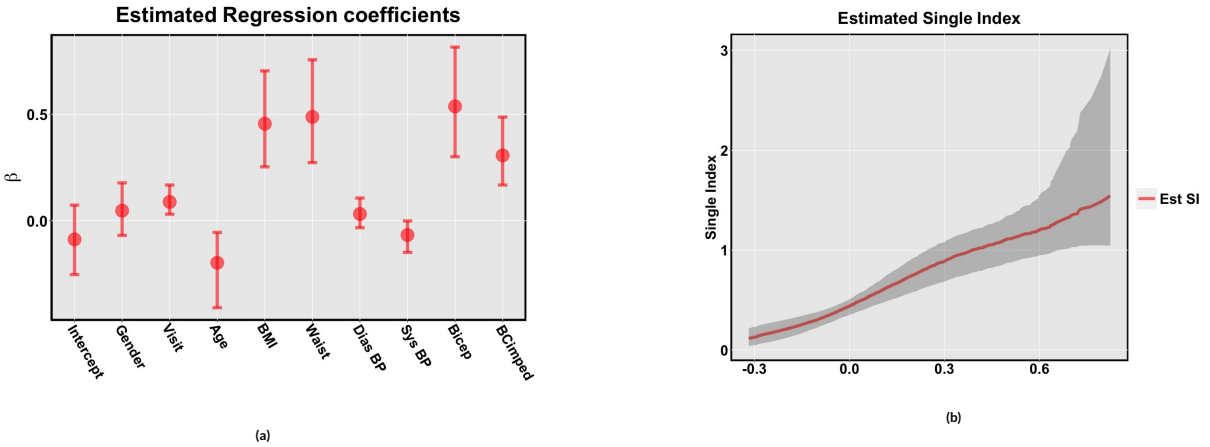

FIGURE 2 *The estimate of regression coefficients (left panel) and single index (right panel) with 95% credible obtained from BR-MSIM*(μ_{it}, ψ).

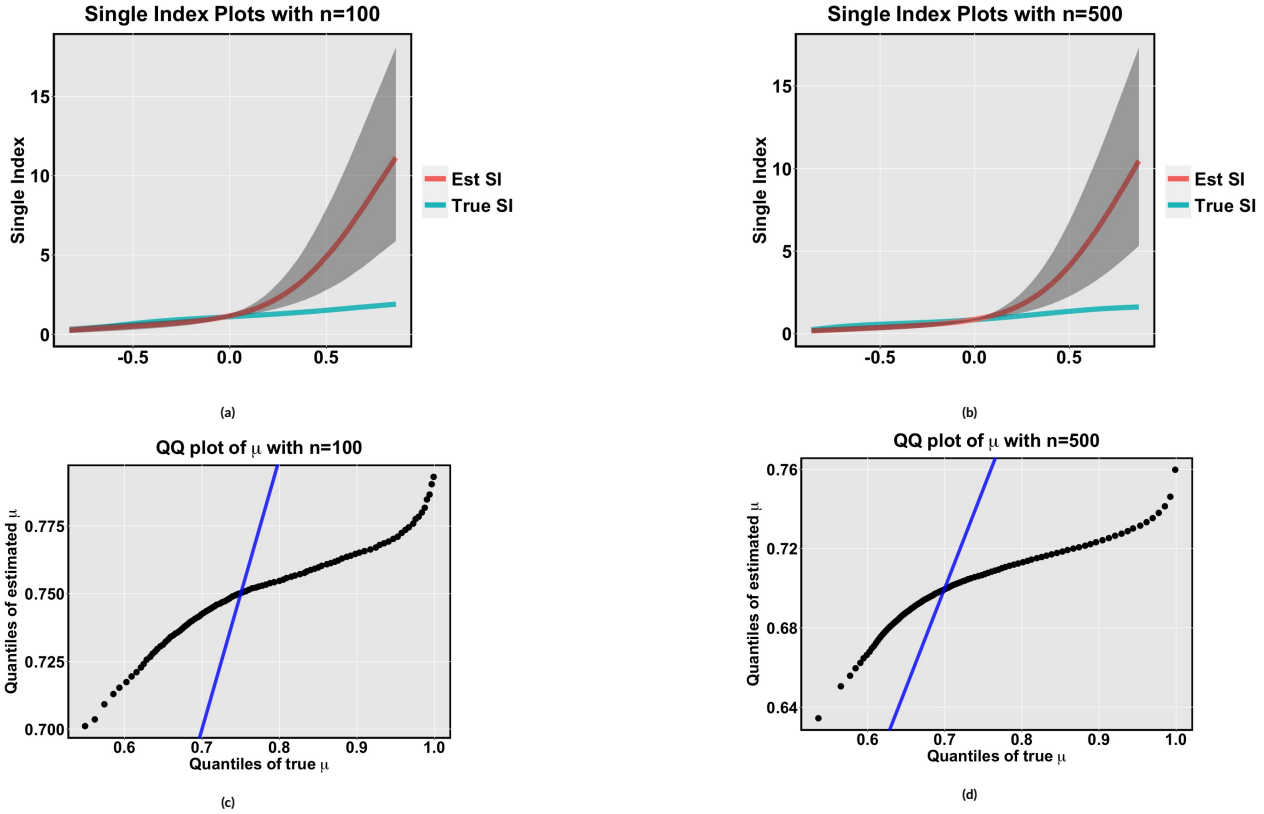

FIGURE 3 *Figure 3a and Figure 3b are overlaid plots true and estimated single index with* 95% *credible interval for n=100 and* $n = 500$ *respectively. We provide the quantile plot of observed vs estimated mean* (μ) *with* $n = 100$ *in Figure 3c and n=500 in Figure 3d.*

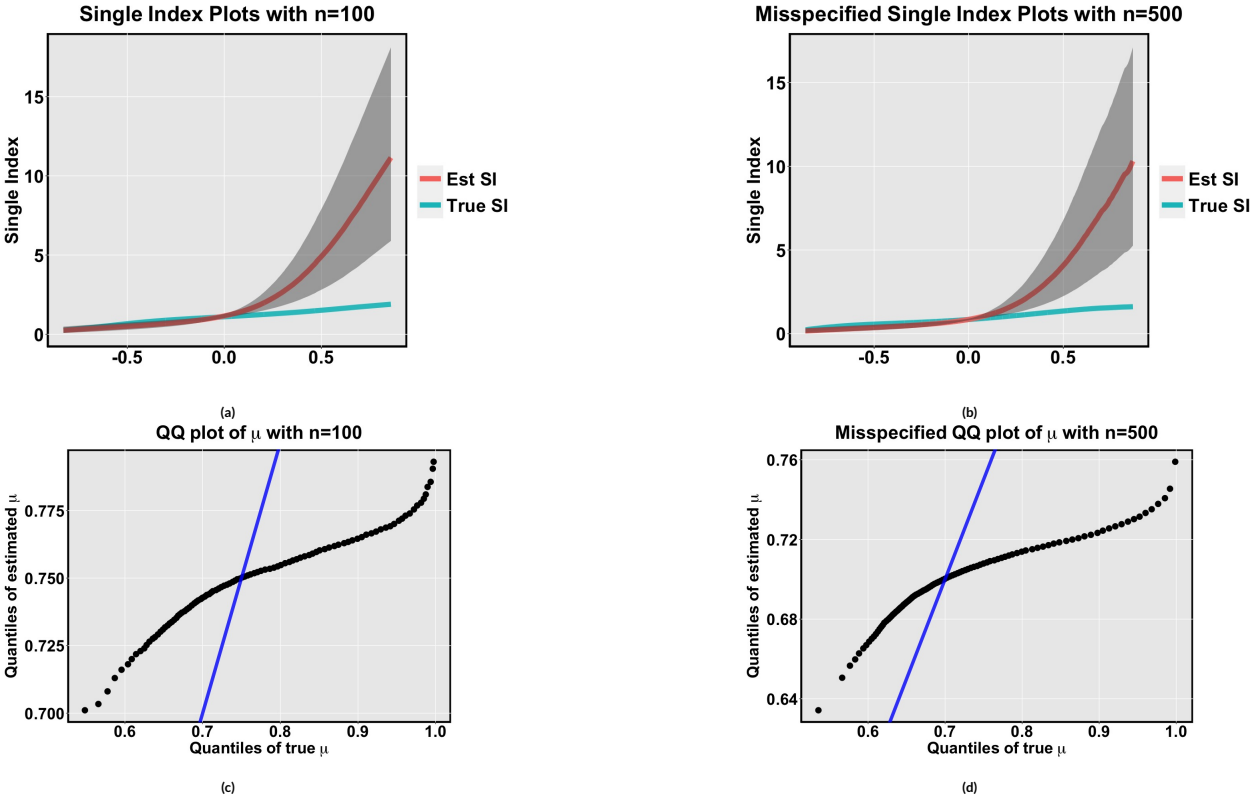

FIGURE 4 *In Figure 4a and 4b, we provide the plot of single index curves for* $n = 100$ *and* $n = 500$ *respectively while the generating data distribution (i.e. a mixture distribution of beta and simplex) is misspecified. Figure 4c and 4d show quantile plot of mean of response variable for* $n = 100$ *and* $n = 500$ *respectively in misspecified case.*

TABLES

	Model 1	Model 2	Model 3	Model 4
WAIC	−7262.58	−7341.75	−12978.72	−13293.56
LPML	3595.38	3639.56	6342.53	6490.95
SSE	16.85	17.06	7.25	6.57
SAE	211.21	214.19	129.87	122.74
MAPE	41.88	42.64	22.03	21.05
MAAPE	0.25	0.25	0.15	0.14

TABLE 1 *Model comparison with WAIC and LPML values of the 4 models.*

	n=100			n=500		
	Bias	MSE	95% CI	Bias	MSE	95% CI
β_1	0.24	0.42	(0.01, 0.69)	0.13	0.31	(0.11, 0.51)
β_2	−0.27	0.34	(0.01, 0.68)	0.20	0.25	(0.16, 0.51)
β_3	−0.04	0.40	(0.02, 0.69)	0.03	0.29	(0.13, 0.60)
β_4	−0.16	0.38	(0.01, 0.69)	−0.09	0.30	(0.10, 0.51)

TABLE 2 *Bias, MSE and* 95% *credible intervals regression coefficients* β *obtained from* (2)*. All the reported values are averaged over* 50 *replicates.*