Bayesian Analysis (2020) **15**, Number 3, pp. 759–780

A New Bayesian Single Index Model with or without Covariates Missing at Random

Kumaresh Dhara[*], Stuart Lipsitz[†], Debdeep Pati[‡], and Debajyoti Sinha[§]

Abstract. For many biomedical, environmental, and economic studies, the single index model provides a practical dimension reaction as well as a good physical interpretation of the unknown nonlinear relationship between the response and its multiple predictors. However, widespread uses of existing Bayesian analysis for such models are lacking in practice due to some major impediments, including slow mixing of the Markov Chain Monte Carlo (MCMC), the inability to deal with missing covariates and a lack of theoretical justification of the rate of convergence of Bayesian estimates. We present a new Bayesian single index model with an associated MCMC algorithm that incorporates an efficient Metropolis–Hastings (MH) step for the conditional distribution of the index vector. Our method leads to a model with good interpretations and prediction, implementable Bayesian inference, fast convergence of the MCMC and a first-time extension to accommodate missing covariates. We also obtain, for the first time, the set of sufficient conditions for obtaining the optimal rate of posterior convergence of the overall regression function. We illustrate the practical advantages of our method and computational tool via reanalysis of an environmental study.

MSC 2010 subject classifications: Primary 62H12; secondary 62G08.

Keywords: Markov Chain Monte Carlo, missing covariates, Gaussian process, mode aligned proposal density, importance sampling.

1 Introduction

For many practical studies, including the environmental study (Chambers, 1983) of Ozone concentration (response) with meteorological covariates, the popular linear model-based analysis is inadequate for inference and prediction when the usual assumptions of linear regression model fail. The Single Index Model (SIM) (Stoker, 1986) provides a simple and interpretable framework for understanding a complex, nonlinear relationship between a response variable Y_i and its $p > 1$ dimensional covariate vector $X_i = (X_{i1}, \cdots, X_{ip})$. The conditional expectation of Y_i given X_i under the SIM is only an unknown univariate function of the scalar index $Z_i = \alpha^{\mathrm{T}} X_i = \sum_{j=1}^{p} \alpha_j X_{ij}$, which is an unknown linear projection of the covariate vector X_i with unknown index vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_p)$. A SIM clearly offers a practical compromise between a completely nonparametric multiple regression and a fully parametric linear regression because it accommodates both nonlinear main effects and higher-order interactions. It also offers

[*] University of Florida, k.dhara@stat.fsu.edu
[†] Harvard Medical School, slipsitz@partners.org
[‡] Texas A&M University, debdeep@stat.tamu.edu
[§] Florida State University, sinhad@stat.fsu.edu

https://doi.org/10.1214/19-BA1170

clear physical interpretations of the index $Z = \alpha^{\mathrm{T}} X$ and relative importance of the predictor effects via the magnitudes of the index parameters $(\alpha_1, \cdots, \alpha_p)$. These practical features of SIM are highly desirable for some biomedical, environmental and econometric studies dealing with the unknown nonlinear relationship between the response and predictors.

There are some great reviews on the recent rapid development of the frequentist SIM literature (Antoniadis et al., 2004; Hristache et al., 2001). These existing methods are essentially of three categories: (a) average derivative method (Stoker, 1986; Powell et al., 1989), (b) M-estimation (Ichimura, 1993; Hardle et al., 1993; Xia et al., 2002; Xia, 2006) and (c) sliced inverse regression (Li and Duan, 1989; Li, 1991). Average derivative methods require highly restrictive and hard to verify (in practice) conditions to ensure the consistency of the estimator of α. M-estimation approaches usually have good asymptotic properties. However, these methods often lead to difficult high-dimensional optimization. The sliced inverse regression methods have only limited applications in practice since they require the distributions of X_i to be elliptically symmetric.

In existing frequentist literature on SIM, the overwhelming emphasis has been on good empirical and asymptotic properties of the point estimates of the index parameter α and the link function $f(\cdot)$. The usual frequentist approaches to characterize uncertainty are based on either an asymptotic approximation or computationally intensive resampling methods. These methods can break down in practice even with a moderate number of covariates. However, Bayesian methods provide a realistic evaluation of the uncertainty of the estimates as well as of the predictions of future responses for many biomedical and other applications. Most of the existing Bayesian approaches for SIM use some basis representation such as splines (Antoniadis et al., 2004; Wang, 2009) and wavelets (Park et al., 2005) of $f(\cdot)$, along with a multivariate prior density on the coefficients of the chosen basis representation of $f(\cdot)$. For these methods, selections of the number of basis functions and the location of the knots for $f(\cdot)$ are subjective and pose computational difficulties. Even reversible jump Markov Chain Monte Carlo algorithms involving movable knots (Wang, 2009) suffer from computationally expensive variable dimensional iterations. Alternatives to the methods with basis representation of f include methods using a Gaussian process prior (Choi et al., 2011; Gramacy and Lian, 2012) on $f(\cdot)$. Each iteration of the Markov Chain Monte Carlo (MCMC) algorithm for such methods needs the inversion of a new $(n \times n)$ dimensional covariance matrix for $f(z_i)$, since the kernel matrix is a function of the index vector α, where $z_i = \alpha^{\mathrm{T}} x_i$ for $i = 1, 2, \cdots, n$. This makes the algorithm computationally intensive even when the sample size n is moderately large. In Section 2, we propose the new Bayesian SIM using the Ornstein–Uhlenbeck (OU) process prior $f(\cdot)$ with a covariance function that helps us avoid the numerical inversion of the $(n \times n)$ covariance matrix within MCMC.

In Section 3, we present a novel and convenient method to generate from the posterior distributions of the parameters of the SIM. For all existing Bayesian methods, a major challenge is to generate samples from the conditional posterior distribution of the index vector α. In general, for the Metropolis–Hastings (MH) step to simulate α within each MCMC iteration, no obvious choice of a proposal density can ensure reasonable acceptance rate and autocorrelation of the posterior samples of α. In this paper, we devise a proposal density based on aligning the mode of the proposal density with that of the target conditional posterior density of α. Our proposal density facilitates a much higher acceptance rate of the MH step for simulating α compared to the acceptance rates for existing methods.

In Section 4, we also present a theoretical justification for our Bayesian method. We provide the minimal regularity conditions required to ensure the optimal convergence rate of the estimate of the overall mean regression function, $g(x) = f(\alpha^{\mathrm{T}} x)$. We provide these conditions for the OU process prior (recommended by us) as well as for the Gaussian process (GP) prior with square exponential covariance kernel. These minimal regularity conditions about our GP prior on the unknown $f(\cdot)$ are easy to ensure in practice and help us determine the prior for $f(\cdot)$ to achieve the optimal convergence rate of the Bayesian estimates of mean response.

Like numerous other biomedical studies, our environmental study example has observations with some missing covariates. To the best of our knowledge, there are very few existing frequentist methods and no Bayesian methods for SIM that accommodate missing covariates. These inverse probability weighted, estimating equations-based frequentist approaches (Xue, 2013; Guo et al., 2015; Li and Yang, 2016) essentially do not use observations with missing covariates. Their other major weaknesses include a difficult to estimate finite sample variance of the estimated α even while using bootstrap, high dependence of the performance of the estimates of $f(\cdot)$ on the choice of the bandwidth of the kernel smoothing and computational difficulty associated with the empirical likelihood method. Related important work of Niu et al. (2017), however, focuses only on statistical tests for the parametric single index model. To address these weaknesses, in Section 5, we develop a novel and computationally efficient Bayesian SIM approach to handle Missing-At-Random (MAR) covariates using post-MCMC importance sampling weights. In Section 6, we present simulation studies to demonstrate the finite sample properties of our Bayesian methods and their comparisons with estimates from some competing methods for whom the software is publicly available.

In Section 7, we present the reanalysis of air quality study (Chambers, 1983) to illustrate the practical advantages of our Bayesian method compared to other existing competitors. Section 8 presents some concluding remarks and discussions about various further extensions and relationships among different approaches.

2 Bayesian Method for SIM

We denote observed data as $D_n = \{(y_i, x_i) : i = 1, 2, \cdots, n\}$, where (y_i, x_i) is the observed values of the response $Y_i \in \mathbb{R}$ and corresponding p-dimensional predictors $X_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$ independent subjects. Without loss of generality, we assume that the covariates are scaled so that $X_i \in [0, 1]^p$. We assume the single-index model

$$Y_i = f(\alpha^{\mathrm{T}} X_i) + \epsilon_i \ , \tag{2.1}$$

with independent errors $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, unknown index parameter $\alpha = (\alpha_1, \cdots, \alpha_p)$ and an unknown nonlinear univariate link-function $f(\cdot)$. Some generalizations of SIM such as projection pursuit regression (PPR) (Friedman et al., 2001) aim to find appropriate multiple projections and link functions of multidimensional X_i to model Y_i. However, the models obtained from PPR are hard to interpret because each component of X_i can affect $E[Y_i|X_i]$ via multiple link functions. To ensure identifiability of the model in (2.1), we further need $\alpha \in \mathbb{S}^+_{p-1} = \{\alpha : \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_p^2 = 1, \alpha_1 > 0\}$ (Lin and Kulasekera, 2007). This restriction on α poses challenges for the specification of prior of the index vector α and for the subsequent Bayesian posterior computation. To address this challenge, we specify a Π_θ for the one-to-one polar transformation θ (Park et al., 2005), where $\alpha_1 = \sin(\theta_1), \alpha_2 = \sin(\theta_2)\cos(\theta_1), \cdots, \alpha_{p-1} = \sin(\theta_{p-1})\prod_{j=1}^{p-2}\cos(\theta_j), \alpha_p = \prod_{j=1}^{p-1}\cos(\theta_j)$ for $\theta_j \in [0, \pi]$ for all $j = 1, \ldots, p-1$ to ensure $\alpha \in \mathbb{S}^+_{p-1}$. A prior Π_α for α corresponds to a specific prior Π_θ for θ. In spite $\alpha = \alpha(\theta)$ being a function of θ, we sometime suppress this relation in the notation α for the brevity. We assume that the joint prior $\Pi(\Lambda)$ of the parameters $\Lambda = (f, \theta, \sigma, \kappa)$ of our Bayesian single index model of (2.1) is

$$\Pi(\Lambda) = \Pi(f, \theta, \sigma, \kappa) \propto \ \Pi_{f|\alpha(\theta),\kappa} \times \Pi_\theta \times \Pi_\sigma \times \Pi_\kappa \ , \tag{2.2}$$

where $\Pi_{f|\alpha(\theta),\kappa} = \Pi_{f|\theta,\kappa}$ is the conditional nonparametric prior for the link-function $f(\cdot)$, given θ and the hyperparameter κ with corresponding hyper-prior Π_κ. Also, Π_θ and Π_σ are marginal priors for θ and σ, respectively. Conditional on α, we require the prior $\Pi_{f|\alpha,\kappa}$ to be supported on $\mathbb{C}[-\sqrt{p}, \sqrt{p}]$, the space of all continuous functions on $[-\sqrt{p}, \sqrt{p}]$, because a conditional prior on the function space $\{f : t \mapsto f(t), t = \alpha^{\mathrm{T}} x\}$ for a fixed α needs the restriction of the domain $|t| \leq \|\alpha\|\,\|x\| \leq \sqrt{p}$ because $||\alpha|| = 1$. A practical choice for the prior model of a nonparametric link function is $f(\cdot)$, which is the Gaussian process prior $\mathrm{GP}(\mu, c_\kappa)$ indexed by the user-specified prior mean function $\mu : \mathbb{R} \to \mathbb{R}$ and the positive definite covariance kernel $c_\kappa : [-\sqrt{p}, \sqrt{p}] \times [-\sqrt{p}, \sqrt{p}] \to \mathbb{R}$, known except the hyper-parameter κ that controls the smoothness of the realizations (sample paths) of $\mathrm{GP}(\mu, c_\kappa)$. The choice of the covariance kernel c_κ is vital for controlling the sample paths of the $\mathrm{GP}(\mu, c_\kappa)$. We opt for the Ornstein–Uhlenbeck (OU) process with the covariance kernel given by $c_\kappa(t_1, t_2) = e^{-\kappa|t_1 - t_2|}$. The OU process is often a popular and appropriate choice, and this $c_\kappa(t_1, t_2)$ is a special case of the more general Matérn family of covariance kernels

$$c(t_1, t_2) = \tau^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu}\sqrt{\kappa}\,|t_1 - t_2|\right)^\nu K_\nu\left(\sqrt{2\nu}\sqrt{\kappa}\,|t_1 - t_2|\right), \tag{2.3}$$

where $\Gamma(\cdot)$ is the gamma function, K_ν is the modified Bessel function of the second kind, ν and κ are the smoothness and bandwidth parameters and τ controls the signal-to-noise ratio. The OU process corresponds to $(\nu = 0.5, \tau = 1)$ in (2.3). Even though the OU stochastic process used as a prior for $f(\cdot)$ has only one unknown hyper-parameter κ, we use the notation $\Pi_{f|\alpha,\kappa}$ for the prior $f(\cdot)$ because the evaluation of the posterior distribution involves the joint multivariate prior density of $f(\cdot)$ evaluated at a finite number of values which are functions of only (α, κ) (explained in the following section). We use the prior Π_κ for κ to be a discrete uniform on a set $\mathcal{J} = \{a : \Pi_\kappa(a) > 0\}$ of equally spaced grid points in the interval $[\kappa_{\min}, \kappa_{\max}]$. This very convenient and practical prior is capable of approximating any continuous prior density for κ to any acceptable level via appropriately choosing the number and gaps of the grid points in $\mathcal{J}$. For the prior Π_σ of σ^2, we use the Inverse-Gamma prior $\text{IG}(\gamma_1, \gamma_2)$ density with shape γ_1 and scale γ_2 for σ^2. This big enough class can also accommodate any reasonable prior opinion about σ^2, and it is also a convenient prior since it is conjugate under the Bayesian SIM model of (2.1).

3 Sampling from Posterior Distribution

Based on the joint prior $\Pi(\Lambda)$ in (2.2) and SIM model in (2.1), the joint posterior $p(\Lambda|D_n)$ given observed data D_n is

$$p(\Lambda|D_n) \propto \phi_n(\mathbf{y}; \mathbf{f}, \sigma^2 I) \times \Pi(\Lambda) \propto \phi_n(\mathbf{y}; \mathbf{f}, \sigma^2 I)\ \Pi_{f|\alpha,\kappa}\ \Pi_\kappa\ \Pi_\theta\ \Pi_\sigma\ , \tag{3.1}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, $\phi_n(\mathbf{y}; \mu, \Sigma)$ is the n-variate normal density with mean-vector μ, and covariance matrix Σ evaluated at $\mathbf{y}$. It is important to note that the joint posterior in (3.1) involves the unknown link function $f(\cdot)$ only through $\mathbf{f} = (f(t_1), \cdots, f(t_N))$, values of $f(\cdot)$ only at $N \le n$ distinct ordered values $t_1 < \cdots < t_N$ of $\{\alpha^{\mathrm{T}} x_1, \cdots, \alpha^{\mathrm{T}} x_n\}$. The OU process prior for $f(\cdot)$ with unknown inverse bandwidth parameter κ implies that the multivariate prior density $\Pi_{f|\alpha,\kappa}$ is essentially a N-variate normal density $\phi_N(\mathbf{f}; 0, \Sigma_{\theta,\kappa})$ evaluated at the $N \le n$-dimensional vector $\mathbf{f} = (f(t_1), \cdots, f(t_N))$ with known mean $(\mu(t_1), \cdots, \mu(t_N))$ and the covariance matrix $(\Sigma_{\alpha,\kappa})_{j,k} = e^{-\kappa|t_j - t_k|}$. In general, we recommend $\mu(t_j) = 0$ for $j = 1, 2, \cdots, N$. This covariance matrix $\Sigma_{\alpha,\kappa}$ is a function of both α (hence its polar transformation θ) and κ. One distinct advantage of using the OU process over other choices of Gaussian processes is that $\Sigma^{-1}_{\alpha,\kappa}$ is a tridiagonal matrix with closed-form expression

$$\Sigma^{-1}_{\alpha,\kappa} = \begin{bmatrix} s_1 & -q_1 & & & & \\ -q_1 & s_2 & -q_2 & & & \\ & & \ddots & \ddots & \ddots & \\ & & & -q_{N-2} & s_{N-1} & -q_{N-1} \\ & & & & -q_{N-1} & s_N \end{bmatrix},$$

where $(\Sigma^{-1}_{\alpha,\kappa})_{i,j} = 0$ for $j \notin \{i+1, i, i-1\}$, upper and lower subdiagonals are $q_i = r_i/(1 - r_i^2)$ with $r_i = exp\{-\kappa(t_{i+1} - t_i)\}$ for $1 \le i \le N-1$, and diagonal entries are $s_i = 1 + r_i q_i + r_{i-1} q_{i-1}$ for $i = 1, \cdots, N$. The determinant d_N of the tridiagonal matrix $\Sigma^{-1}_{\alpha,\kappa}$ (and hence the determinant of $\Sigma_{\alpha,\kappa}$) is also evaluated via the sequential formula $d_m = s_m d_{m-1} - q^2_{m-1} d_{m-2}$ with $d_1 = s_1$ and $d_0 = 1$. These help the implementation of the MCMC algorithm via avoiding the numerical inversion and direct computation of the determinant of any potentially high-dimensional covariance matrix at each iteration. The resulting conditional posteriors for the MCMC algorithm are the following:

$$\mathbf{f} \mid \sigma^2, \kappa, \theta, D_n \sim \phi_n\{\mathbf{f}; (\Sigma^{-1}_{\theta,\kappa} + \sigma^{-2} I_n)^{-1}\mathbf{y}/\sigma^2, (\Sigma^{-1}_{\theta,\kappa} + \sigma^{-2} I_n)^{-1}\}, \tag{3.2}$$

$$\sigma^2 \mid \mathbf{f}, \kappa, \theta, D_n \sim \text{IG}\left(\frac{n+\gamma_1}{2}, \frac{\|\mathbf{y} - \mathbf{f}\|^2 + \gamma_2}{2}\right), \tag{3.3}$$

$$P(\kappa = \kappa^{(t)} \mid \sigma^2, \mathbf{f}, \theta, D_n) \quad = \quad \frac{\exp\big\{-\mathbf{f}'\Sigma^{-1}_{\theta,\kappa^{(t)}}\mathbf{f}/2\big\}}{\sum_{s\in\mathcal{J}}\exp\big\{-\mathbf{f}'\Sigma^{-1}_{\theta,\kappa^{(s)}}\mathbf{f}/2\big\}} \quad \text{for } \ t \in \mathcal{J}, \tag{3.4}$$

$$p(\theta|\sigma^2, \kappa, \mathbf{f}, D_n) \quad \propto \quad |\Sigma_{\theta,\kappa}|^{-1/2} e^{-\mathbf{f}'\Sigma^{-1}_{\theta,\kappa}\mathbf{f}/2}\Pi_\theta(\theta), \tag{3.5}$$

where $\Pi_\theta(\theta)$ is the prior for the one-one polar transformation θ of α (as explained earlier), $\phi_n(.;\mu, B)$ is the n-variate normal density with mean μ and variance matrix B and $\mathcal{J}$ is the discrete support of the prior of κ. For the sake of brevity, the above conditional posteriors are given for the case $N = n$ when all the values of $\alpha^{\mathrm{T}}x_i$ are distinct. We again emphasize that the conditional posterior distributions in (3.2)–(3.4) are straightforward to sample from because we have a closed-form expression for $\Sigma^{-1}_{\theta,\kappa}$ and a sequential formula for its determinant.

However, for any choice of prior $\Pi_\theta(\theta)$, sampling θ from the conditional density of (3.5) is not straightforward in spite of the available expressions for $\Sigma^{-1}_{\theta,\kappa}$ and $|\Sigma_{\theta,\kappa}|^{-1/2}$. Hence, we use a Metropolis–Hastings (MH) step within MCMC iteration to sample θ from the target density in (3.5). For this MH step, our independent proposal densities $\mathcal{P}_j$ of θ_j for $j = 1, \cdots, p-1$ are rescaled Beta densities with common support $[0, \pi]$, that is $\mathcal{P}_j(\theta_j) = \mathcal{B}(\theta_j/\pi; c_j, d_j)/\pi$, where $\mathcal{B}(u; c, d) \propto u^{c-1}(1-u)^{d-1}$ is the Beta density with parameters (c, d). To assure an appropriate proposal density $\mathcal{P}_j(\theta_j)$ with good acceptance rate for the MH step, we choose d_j to satisfy $\hat{\theta}_{j,\mathrm{MAP}} = (\pi c_j d_j)/\{(c_j + d_j)^2(c_j+d_j+1)\}$ for fixed value of c_j, where $\hat{\theta}_{\mathrm{MAP}} = (\hat{\theta}_{1,\mathrm{MAP}}, \cdots, \hat{\theta}_{p,\mathrm{MAP}})$ is the maxima of the target density in (3.5). We call this $\hat{\theta}_{\mathrm{MAP}}$, as the conditional maximum a posteriori (MAP) of θ. As the support of θ in (3.5) is closed and bounded, we initially fix a set of grid points $\mathbb{G}$, and take $\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta_r\in\mathbb{G}}\, p(\theta_r|\sigma^2, \kappa, \mathbf{f}, D)$ where $p(\theta_r|\sigma^2, \kappa, \mathbf{f}, D)$ is the conditional posterior density of θ evaluated at θ_r. The approximation of the mode by $\hat{\theta}_{\mathrm{MAP}}$ is improved by making the grid points in $\mathbb{G}$ finer. An alternative procedure for computing $\hat{\theta}_{\mathrm{MAP}}$ is to use any built-in optimization algorithm available in `R` or `MATLAB`. Increasing the value of c_j results in lowering the variance, if d_j is kept fixed. We decide the value of c_j based on the desired variance of the proposal density. For simplicity, we fix all the values of $c_j = c$ for $j = 1, 2, \cdots, p-1$.

Alternatively, sampling θ from (3.5) and f from (3.2) within each MCMC iteration results in highly autocorrelated posterior samples since f and θ are highly correlated in the posterior. To circumvent this problem, we modify the MH step for sampling θ by adapting the method of Murray and Adams (2010), using surrogate data $\mathbf{h}$. We describe the MCMC algorithm as follows:

A possible choice of S_θ is $\tau^2 I$ where τ^2 is a fixed, say, $\tau^2 = 0.1$. In the above expressions, $R_\theta = (\Sigma_\theta^{-1} + S_\theta^{-1})^{-1}$, $m_{\theta,h} = R_\theta S_\theta^{-1}\mathbf{h}$, Δ_R is the Cholesky decomposition of matrix R with $R = \Delta_R\Delta_R^{\mathrm{T}}$. The transition kernels used for computing the acceptance probability of new $\tilde{\theta}$ based on past sample $\theta^{(t)}$ are $q(\theta^{(t)}; \tilde{\theta}) = \prod_{j=1}^{p-1}\mathcal{P}_j(\theta_j^{(t)}|\tilde{c}_j, \tilde{d}_j)$ and $q(\tilde{\theta}; \theta^{(t)}) = \prod_{j=1}^{p-1}\mathcal{P}_j(\tilde{\theta}_j|c_j^{(t)}, d_j^{(t)})$. Here, $(\tilde{c}_j, \tilde{d}_j)$ are the parameters for the proposal density used for generating $\tilde{\theta}$ and $(c_j^{(t)}, d_j^{(t)})$ are the parameters for the proposal density used in previous iteration number T.

Algorithm 1 Simultaneous update of $(\theta, \mathbf{f})$

```
Start with $(\theta^{(t)}, \mathbf{f}^{(t)})$, at iteration t
Simulate surrogate data $\mathbf{h}$ from $\phi_n(\cdot; \mathbf{f}^{(t)}, S_{\theta^{(t)}})$
Compute the latent variable $\eta = \Delta^{-1}_{R_{\theta^{(t)}}}(\mathbf{f}^{(t)} - m_{\theta^{(t)},h})$
Generate $\tilde{\theta}$ from the proposal density $\mathcal{P}_1 \times \mathcal{P}_2 \times \cdots \mathcal{P}_{p-1}$
Compute the new proposal $\tilde{\mathbf{f}} = m_{\tilde{\theta},h} + \Delta_{R_{\tilde{\theta}}}\eta$
Draw $u \sim \mathrm{U}(0,1)$
If $u < \frac{L(\tilde{\mathbf{f}})\phi_n(h;0,\Sigma_{\tilde{\theta},\kappa}+S_{\tilde{\theta}})\pi_\theta(\tilde{\theta})q(\theta^{(t)};\tilde{\theta})}{L(\mathbf{f}^{(t)})\phi(h;0,\Sigma_{\theta^{(t)},\kappa}+S_{\theta^{(t)}})\pi_\theta(\theta^{(t)})q(\tilde{\theta};\theta^{(t)})}$ then $(\theta^{(t+1)}, \mathbf{f}^{(t+1)}) = (\tilde{\theta}, \tilde{\mathbf{f}})$
Else $(\theta^{(t+1)}, \mathbf{f}^{(t+1)}) = (\theta^{(t)}, \mathbf{f}^{(t)})$
```

4 Theoretical Results

The semiparametric Bayesian estimator of the regression function $g(x) = f(\alpha^{\mathrm{T}}x)$ in Choi et al. (2011) has been shown to possess posterior consistency. However, the theory and associated conditions for obtaining a desirable convergence rate of the posterior estimate of $g(x)$ are not available yet. Even though we do not evaluate the convergence rates of the Bayesian estimates of $f(\cdot)$ and α separately, our theoretical result describes how the convergence rate of the Bayesian estimate of regression $g(x)$ to the true regression function $g_0(x)$ depends on the sample size n and the roughness β of the true link function $f_0(\cdot)$. This result also shows that the asymptotic performance of our Bayesian estimate of $g(x)$ at the observed covariate values $x_1, \ldots, x_n$ is optimal as long as the roughness of the sample path of the prior process of the unknown $f(\cdot)$ is not too high compared to the sample-size n. We also show that this desirable result holds even when the roughness β of the true $f_0(\cdot)$ is not very high (i.e., f_0 is less than twice differentiable).

Let $\mathbb{C}[0,1]^p$ and $\mathbb{C}^\beta[0,1]^p$, respectively, denote the space of continuous functions and the Hölder space of β-smooth functions $f : [0,1]^p \to \mathbb{R}$, endowed with the supremum norm $\|f\|_\infty = \sup_{t \in [0,1]^p} |f(t)|$. For $\beta > 0$, the Hölder space $\mathbb{C}^\beta[0,1]^p$ consists of functions $f \in \mathbb{C}[0,1]^p$ that have bounded mixed partial derivatives up to order $\lfloor \beta \rfloor$, with the partial derivatives of order $\lfloor \beta \rfloor$ being Lipschitz continuous of order $\beta - \lfloor \beta \rfloor$. The Sobolev space $\mathbb{H}^\beta[0,1]^p$ is the set of functions $f : [0,1]^p \to \mathbb{R}$ that are restrictions of a function $f : \mathbb{R}^p \to \mathbb{R}$ with Fourier transform $\hat{f}(\lambda)$ satisfying $\|f\|^2_\beta := \int (1+\|\lambda\|^2)^\beta \left|\hat{f}(\lambda)\right|^2 d\lambda < \infty$ where $\|f\|_\beta$ is the Sobolev norm of f. Roughly speaking, for integer β, a function belongs to $\mathbb{H}^\beta$ if it has partial derivatives up to order β that are all square-integrable.

For the sake of brevity, we consider a special case of the Bayesian SIM model described in (2.1) with $\sigma^2 = 1$ for our theoretical development. Let the true data generating parameters be $(f_0(\cdot), \alpha_0)$. We now state the main result on posterior contraction in estimating $g_0(x) = f_0(\alpha_0^{\mathrm{T}}x)$ with respect to the empirical L_2 norm given by $\|g - g_0\|^2_{2,n} = (1/n)\sum_{i=1}^n \{g(x_i) - g_0(x_i)\}^2$. In particular, we would like to find the minimum possible sequence of positive numbers ϵ_n converging to 0 such that

$$\mathbb{E}_0\mathbb{P}(\|g - g_0\|_{2,n} \leq M\epsilon_n \mid D_n) \to 1 \tag{4.1}$$

for a suitably large positive number M, where $\mathbb{E}_0$ denotes expectation under the true data generation mechanism and $D_n = (\mathbf{y}, X)$ is the observed data indexed by sample size n. Assuming $g_0 \in \mathbb{C}^\beta[0,1]^p$ and recognizing that g_0 is essentially concentrated on a subspace with dimension 1, the optimal (in a minimax sense) rate for estimating g_0 is much faster ($\epsilon_n = n^{-\beta/(2\beta+1)}$) instead of $\epsilon_n = n^{-\beta/(2\beta+p)}$, the minimax rate of estimating a p-variable function with smoothness β. We investigate the concentration rates for two different choices of prior on the link function f_0. Theorem 1 presents the concentration rate when the prior density on f_0 is a OU process, i.e., the $(i,j)^{th}$ element of the covariance matrix is $c_{ij} = e^{-\kappa|\alpha^{\mathrm{T}} x_i - \alpha^{\mathrm{T}} x_j|}$. Theorem 2 provides the concentration rate when the prior density on f_0 is a Gaussian process with square exponential covariance kernel, i.e., the $(i,j)^{th}$ element of the covariance matrix is $c_{ij} = e^{-\kappa(\alpha^{\mathrm{T}} x_i - \alpha^{\mathrm{T}} x_j)^2}$.

Theorem 1. *If $f_0 \in \mathbb{C}^\beta[0,1] \cap \mathbb{H}^\beta[0,1]$ and prior on f_0 is an OU process, then (4.1) is satisfied with $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ for some $t > 0$ and for OU process inverse bandwidth parameter $\sqrt{\kappa} \equiv n^{(1-2\beta)/(2\beta+1)}$.*

Theorem 2. *If $f_0 \in \mathbb{C}^\beta[0,1]$ and prior on f_0 is a Gaussian process with a square exponential covariance kernel, then (4.1) is satisfied with $\epsilon_n = n^{-\beta/(2\beta+1)}(\log n)^t$ for some $t > 0$ and for inverse bandwidth parameter $\sqrt{\kappa} \equiv n^{1/(2\beta+1)}$.*

Theorem 1 shows that we achieve the optimal rate of convergence of the posterior around true mean $g_0(x)$ if the square root of inverse bandwidth parameter $\sqrt{\kappa}$ of the OU process prior is taken to be of order $n^{(1-2\beta)/(2\beta+1)}$. It is interesting to note that for $\beta < 1/2$ (i.e., when the true function f_0 is not even differentiable), we need κ of the OU process prior to go to infinity to achieve the optimal rate of convergence. However, when $\beta > 1/2$, we need κ to go to 0 to ensure the optimal convergence rate. Such conditions are required since the OU process has extremely rough sample paths. However, for the Gaussian prior, Theorem 2 shows that irrespective of the smoothness of f_0, we need to ensure that κ of the Gaussian process goes to infinity to ensure the optimal concentration rate of the posterior. This is expected since the sample paths of a Gaussian process with square exponential covariance kernel are smooth. Therefore, to ensure that the posterior estimate of f converges to f_0 at an optimal rate, we need to ensure that κ goes to infinity. The proofs of Theorem 1 and 2 are provided in Supplementary Materials (Appendix B, Dhara et al., 2019).

5 Analysis with Missing-At-Random Covariates

To the best of our knowledge, our paper presents the first extension of the Bayesian single-index model to accommodate covariates that are missing at random (Little and Rubin, 2014). By an abuse of notation, the entire data D_n is expressed as $D_n = D_1 \cup D_2$, where D_1 and D_2 contain all the observed data from subjects respectively in S_1 and S_2. Here, S_1 and S_2 are respectively the set of subjects with no missing covariates and the set of subjects with at least 1 missing covariate. We also use the notation $X_{ci} = (X_{im}, x_{io})$ to denote the "complete" covariate vector from each subject $i \in S_2$, where X_{im} denotes the unknown missing covariates and x_{io} denotes the observed parts of the covariates. For each subject $i \in S_1$, we have $X_{c_i} = x_{i_0}$ with X_{im} being an empty vector. We assume that X_{ci} has the joint density $g_X(\cdot \mid \gamma)$ independently for $i = 1, \cdots, n$. It is reasonable and conventional in missing-data literature (Little and Rubin, 2014) to assume that γ shares no parameters with the set of parameters $\Lambda = (f, \theta, \sigma, \kappa)$ associated with the regression model (2.1) and its prior Π_γ is independent of the joint prior $\Pi(\Lambda)$ in (2.2).

Now, the joint posterior based on the entire observed data D_n is

$$p(\Lambda, \gamma \mid D_n) \propto L_1(D_1 \mid \Lambda, \gamma) \times L_2(D_2 \mid \Lambda, \gamma) \times \Pi(\Lambda) \times \Pi_\gamma(\gamma), \tag{5.1}$$

where the likelihood based on D_1 is

$$L_1(D_1 \mid \Lambda) \propto \prod_{i \in S_1} \phi(y_i - f(\alpha^{\mathrm{T}} x_{io}); 0, \sigma)\ g_X(x_{io} \mid \gamma)\ ,$$

the likelihood based on D_2 is

$$L_2(D_2 \mid \Lambda) \propto \prod_{i \in S_2} \int \{\phi(y_i - f(\alpha^{\mathrm{T}} X_{ic}); 0, \sigma)\ g_X(X_{im}, x_{io} \mid \gamma)\}\ dX_{im}\ , \tag{5.2}$$

and $\Pi_\gamma(\gamma)$ is the prior distribution of γ based on the fully observed covariates x_{io} for the subjects $i \in S_1$. For L_2 in (5.2), each integral is computed over the sample space of all missing covariates of the subject $i \in S_2$. For example, in our analysis of the air-quality study, we assume $g_X(\cdot \mid \gamma)$ to be the multivariate Gaussian density $\phi_p(x; \mu, \Sigma)$ with the popular and conjugate Normal-inverse-Wishart prior $\pi(\gamma)$ for $\gamma = (\mu, \Sigma)$, even though other multivariate distributions can also be accommodated similarly. In this case, the integration of (5.2) has a closed-form expression. Bayesian estimates from (5.1) need Monte Carlo integration with respect to the joint posterior of $p(\Lambda, \gamma \mid D_n)$. To address this challenge, we propose the following method weighted Mote Carlo integration. Note that equation (5.1) can also be expressed as

$$p(\Lambda, \gamma \mid D_n) \propto p(\Lambda | D_1) L_2(D_2 | \lambda, \gamma) p_1(\gamma | D_1), \tag{5.3}$$

where $p(\Lambda|D_1)$ is the posterior of (3.1) based on D_1 instead of the whole data D_n; $p_1(\gamma|D_1) \propto \{\prod_{i \in S_1} \phi_p(x_i|\gamma)\}\Pi_\gamma(\gamma)$ is the posterior of γ based on fully observed covariates x_{io} from D_1. For the weighted Monte Carlo integration using weighted samples (instead of usual identically distributed samples from standard MCMC) (Λ^*, Γ^*), we first obtain Λ^* from $p(\Lambda|D_1)$ using the MCMC method described in Section 3. Then we independently sample γ^* from the posterior density $\Pi_\gamma(\gamma \mid D_1)$. For example, using the usual conjugate Normal-inverse-Wishart prior $\Pi(\mu, \Sigma)$ of $\gamma = (\mu, \Sigma)$ for our analysis, this $p_1(\mu, \Sigma | D_1)$ has the corresponding Normal-inverse-Wishart posterior density. We then compute the sampling weight w^* of (Λ^*, γ^*) to be proportional to $L_2(D_2 \mid \Lambda^*)$ using integration of X_{im} for all subjects $i \in S_2$. We note that, in our case, with the multivariate normal density for $g_X(X_{im}, x_{io}|\gamma)$, the X_{im} given (x_{i0}, γ^*) has the multivariate normal distribution with a mean and covariance matrix that are known functions of x_{io} and γ^*. Unlike our case of multivariate normal density for $g_X(X_{im}, x_{io}|\gamma)$, one may need to perform a Monte Carlo integration for $L_2(D_2 \mid \Lambda^*)$ via generating the missing observations X_{im} from the conditional density of X_{im} given (x_{i0}, γ^*). Finally, to obtain the Bayesian estimates, we use the samples (Λ^*, Γ^*) with the corresponding sampling weights w^*.

6 Simulation Studies

Throughout our simulation studies, we use the OU process $\Pi_{f|\alpha,\kappa}$ as a prior of the link function $f(\cdot)$, with hyper-parameter κ assigned a discrete uniform prior on the interval $[0.5, 2]$ with grid size 0.05. The prior distribution for error variance σ^2 is inverse gamma with shape parameter 2 and rate parameter 0.01. For evaluating the approximate mode of the conditional posterior of θ within MH steps, we use grid size 0.05 for each θ_j. We fix each c_j at 5,000. We estimated d_j using the method described in Section 3. For each Bayesian analysis based on our and other competing Bayesian methods, we use 3,000 MCMC samples after discarding the first 2,000 for burn-in.

In our simulation study 1, we investigate the performance of our Bayesian estimates based on the SIM of (2.1) (BSIM in short) for different sample sizes ($n = 50, 100$), forms of true link functions and values of the error standard deviations $\sigma \in \{0.01, 0.5\}$. Two different true link functions f_0 considered here are quadratic with $f_1(z) = z + z^2$ and exponential with $f_2(z) = e^z$. True index parameter $\alpha_0 = (1, 1, 1)/\sqrt{3}$ corresponds to true $\theta_0 = (0.61547, 0.7854)$ in polar coordinates. We use 50 replicates of the datasets for each combination of true f_0, σ and n to approximate the sampling distribution of the estimates. We begin with the performance of the estimates when the true data generating simulation models are (2.1) with $\mathrm{N}(0, \sigma^2)$ error distributions. The approximate sampling coverage probability of interval estimates and mean acceptance rate of the Metropolis–Hastings (MH) step are reported in Table 1.

n	σ	True f	Coverage for θ_1	Coverage for θ_2	Acceptance Rate
50	0.01	Exponential	98%	98%	23.52%
	0.01	Quadratic	100%	100%	19.73%
	0.1	Exponential	98%	98%	16.12%
	0.1	Quadratic	100%	94%	16.38%
	0.5	Exponential	70%	78%	16.13%
	0.5	Quadratic	100%	94%	16.85%
100	0.01	Exponential	100%	100%	16.41%
	0.01	Quadratic	98%	96%	16.39%
	0.1	Exponential	100%	98%	16.63%
	0.1	Quadratic	100%	98%	16.07%
	0.5	Exponential	88%	86%	15.83%
	0.5	Quadratic	94%	64%	16.95%

Table 1: Approximate coverage probabilities of BSIM-based interval estimates and acceptance rates for the MH step of $\theta = (\theta_1, \theta_2)$ for different sample sizes n, error standard deviation σ and forms of true link f, assuming that the true model is SIM. The (approximate) coverage probabilities are based on 50 replications of the simulated datasets. The acceptance rate of the MH algorithm is the proportion of times (out of final 3,000 iterations of MCMC) the MH step of θ accepts a new value.

For both forms of true link function f, the average MH acceptance rates of θ are close to 16%. The approximate coverage probabilities of the interval estimates are smaller when the true link function is exponential (f_2) compared to when it is quadratic (f_1). As expected, an increase in error standard deviation σ also lowers the coverage probabilities irrespective of the form of true $f(\cdot)$. The coverage probability is substantially smaller because the signal-to-noise ratio (the standard deviation of $f(Z)$ divided by σ) when $\sigma = 0.5$ is $1/5$ times the signal-to-noise ratio when $\sigma = 0.1$.

Next, in simulation study 2, we compare our BSIM estimates of θ with the estimates from two competing methods: (1) method of Choi et al. (2011) (referred to as Choi-SIM henceforth) and (2) method of first-step projection pursuit regression (PPR). For brevity, we provide the details of simulation study 2 in the Supplementary Materials (Appendix A, Dhara et al., 2019). We have shown that the proposed BSIM method is superior to both Choi-SIM and PPR when the true model is the SIM of (2.1).

In simulation study 3, we compare our BSIM method with two the frequentist methods implemented in the R package `MAVE`, namely, (1) the kernel-based sliced inverse regression (KSIR in short) method of Li (1991) and (2) the method of central dimension reduction using outer product gradient (CSOPG in short) of Xia (2006). We also compare our method with two competing Bayesian methods, both implemented using MCMC methods, namely (1) the Bayesian method of Antoniadis et al. (2004) abbreviated as Antoniadis-SIM, and (2) the Bayesian method of Gramacy and Lian (2012) abbreviated as Gramacy-SIM (R package `tgp`). In order to have a fair comparison among Bayesian and frequentist methods for different n, σ and forms of true f, we compare the competing methods using the median sum of square errors (based on 50 replicates) of the estimates $\hat{\alpha}$ of index vector α. MCMC chains for Gramacy-SIM had a problem of slow mixing of the MCMC and needed a reasonable estimate of the covariance matrix for α to ensure good mixing of MCMC. So, we have used a pilot run with a small number of iterations to compute the variance matrix for α and then used that estimate to run the final set of MCMC. For a fair comparison, we also normalized the final estimates $\hat{\alpha}$ from Gramacy-SIM because of the norm of the initial $\hat{\alpha}$ from Gramacy-SIM may not be 1. Based on the MSSE values from different methods presented in Table 2 for different choices of n and σ, BSIM-based estimates perform similar to CSOPG and Antoniadis-SIM based estimates when σ is small. For larger σ, CSOPG and Antoniadis-SIM perform as well as BSIM. Even though KSIR and Gramacy-SIM perform similar to BSIM when σ is smaller, for higher values of σ, BSIM performs better than both of them.

In simulation study 4, we also compare the MSSE of $\hat{\alpha}$ from our BSIM with those from competing methods when the true underlying distribution of error ϵ_i is a mixture of two normal densities $0.95\ N(0, \sigma^2) + 0.05\ N(0, (3\sigma)^2)$ instead of the normal distribution assumed for SIM. In Table 3, the MSSE values from different methods are presented for two sample sizes and 3 different values of σ. Similar to previous simulation study 3, $\hat{\alpha}$ from BSIM performs better than those from KSIR and Gramacy-SIM. When σ is smaller, BSIM performs better than CSOPG and Antoniadis-SIM. However, for a larger value of σ, BSIM does not outperform CSOPG and Antoniadis-SIM as far the MSSE of $\hat{\alpha}$ is concerned.

Our simulation study 5 investigates the performance of the extension of our BSIM method to deal with missing-at-random (MAR) covariates, under link functions f_1 and f_2, sample sizes $n = 50$ and 100 and error standard deviations $\sigma \in \{0.01, 0.1, 0.5\}$.

n	σ	True f	CSOPG	KSIR	BSIM	Antoniadis-SIM	Gramacy-SIM
50	0.01	Exponential	0.0004	0.0016	0.0004	0.0014	0.0079
	0.01	Quadratic	0.0009	0.0718	0.0003	0.0007	0.0079
	0.1	Exponential	0.0010	0.0019	0.0011	0.0010	0.0092
	0.1	Quadratic	0.0013	0.0795	0.0005	0.0013	0.0091
	0.3	Exponential	0.0045	0.0053	0.0053	0.0023	0.0086
	0.3	Quadratic	0.0022	0.0894	0.0043	0.0053	0.0084
100	0.01	Exponential	0.0001	0.0004	0.0002	0.0006	0.0082
	0.01	Quadratic	0.0002	0.0387	0.0002	0.0005	0.0080
	0.1	Exponential	0.0003	0.0006	0.0006	0.0006	0.0086
	0.1	Quadratic	0.0004	0.0424	0.0007	0.0008	0.0088
	0.3	Exponential	0.0015	0.0019	0.0068	0.0014	0.0082
	0.3	Quadratic	0.0014	0.0352	0.0067	0.0021	0.0087

Table 2: Comparison of MSSE of estimates of α from CSOPG, KSIR, BSIM, Antoniadis-SIM and Gramacy-SIM methods for different sample sizes, true link functions f and error standard deviations σ.

n	σ	True f	CSOPG	KSIR	BSIM	Antoniadis-SIM	Gramacy-SIM
50	0.01	Exponential	0.0005	0.0015	0.0003	0.0015	0.0080
	0.01	Quadratic	0.0009	0.0657	0.0003	0.0015	0.0080
	0.1	Exponential	0.0010	0.0019	0.0012	0.0012	0.0086
	0.1	Quadratic	0.0013	0.0921	0.0008	0.0015	0.0077
	0.3	Exponential	0.0056	0.0071	0.0084	0.0033	0.0114
	0.3	Quadratic	0.0022	0.0868	0.0052	0.0071	0.0088
100	0.01	Exponential	0.0001	0.0003	0.0002	0.0006	0.0082
	0.01	Quadratic	0.0002	0.0407	0.0002	0.0009	0.0080
	0.1	Exponential	0.0004	0.0008	0.0007	0.0007	0.0092
	0.1	Quadratic	0.0004	0.0378	0.0008	0.0014	0.0083
	0.3	Exponential	0.0015	0.0029	0.0089	0.0013	0.0084
	0.3	Quadratic	0.0012	0.0448	0.0077	0.0031	0.0084

Table 3: Comparison of the median of the sum of square error (MSSE) of $\hat{\alpha}$ from CSOPG, KSIR, BSIM, Antoniadis-SIM and Gramacy-SIM methods with the true error distribution 0.95N(0, σ^2)+0.05N(0, $(3\sigma)^2$).

Each of the 50 replicated datasets for each combination of (f, n, σ) has only variable X_1 potentially missing-at-random (MAR) with probability p_{1j} that does not depend on X_1. For the results in Table 4, we use $p_{1j} = 1/(1 + e^{4+2x_{2i}+3x_{3i}+y_i})$, which results in about 23% of observations with X_1 MAR. For the results in Table 5, we use $p_{1j} = 1/(1+e^{2+2x_{2i}+3x_{3i}+y_i})$ which results in approximately 40% observations with X_1 MAR. In these two tables, we compare the BSIM method using only completely observed data D_1 with our proposed BSIM extension for MAR covariates using full data $D = D_1 \cup D_2$. The comparisons of these two competing Bayesian methods are based on the average widths of the Bayesian interval estimates of θ parameters. Not unexpectedly, we obtain narrower 95% credible intervals when we use BSIM extension for available data $D = D_1 \cup D_2 =$ instead of using BSIM only on D_1. These two tables also present the percentage of improvement in average widths of the interval estimates for using BSIM extension for

			Length of Interval				Percentage	
			Using only D_1		Using $D_1 \cup D_2$		Improvement	
n	σ	True f	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
50	0.01	Exponential	0.0889	0.0945	0.0845	0.0939	1.98%	1.35%
	0.01	Quadratic	0.0800	0.0859	0.0784	0.0834	0.51%	0.89%
	0.1	Exponential	0.1184	0.1180	0.1134	0.1120	2.62%	1.57%
	0.1	Quadratic	0.1027	0.1076	0.1007	0.1078	0.81%	0.24%
	0.5	Exponential	0.4801	0.4194	0.4710	0.4220	1.02%	0.28%
	0.5	Quadratic	0.4020	0.3810	0.3910	0.3754	2.74%	1.47%
100	0.01	Exponential	0.0648	0.0760	0.0609	0.0676	6.35%	6.23%
	0.01	Quadratic	0.0651	0.0731	0.0556	0.0565	13.66%	18.03%
	0.1	Exponential	0.0979	0.1084	0.0932	0.0968	5.28%	5.13%
	0.1	Quadratic	0.0929	0.1062	0.0840	0.0888	11.15%	9.95%
	0.5	Exponential	0.3579	0.3643	0.3045	0.2964	11.98%	7.24%
	0.5	Quadratic	0.3601	0.3141	0.2990	0.2440	15.17%	15.56%

Table 4: Comparison of the average length of the 95% Bayesian credible intervals from the complete part of the data (D_1) with average length of the 95% credible intervals from the whole data $D = D_1 \cup D_2$, where D_2 is the part with missing observations. The percentage improvement in the length of the interval is reported in the last two columns. In this simulation, 20% of the observations have X_1 missing.

			Length of Interval				Percentage	
			Using only D_1		Using $D_1 \cup D_2$		Improvement	
n	σ	True f	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
50	0.01	Exponential	0.0954	0.0989	0.0894	0.0897	6.30%	9.27%
	0.01	Quadratic	0.0940	0.0921	0.0818	0.0799	12.92%	13.25%
	0.1	Exponential	0.1101	0.1115	0.1039	0.1057	5.57%	5.22%
	0.1	Quadratic	0.1111	0.1057	0.1010	0.0956	9.10%	9.62%
	0.5	Exponential	0.4671	0.3573	0.3341	0.2771	28.48%	22.44%
	0.5	Quadratic	0.2498	0.2544	0.2297	0.2197	8.04%	13.65%
100	0.01	Exponential	0.0684	0.0790	0.0504	0.0498	26.34%	36.85%
	0.01	Quadratic	0.0668	0.0739	0.0151	0.0157	77.28%	78.82%
	0.1	Exponential	0.0954	0.1033	0.0779	0.0778	18.37%	24.69%
	0.1	Quadratic	0.0907	0.0990	0.0617	0.0621	31.93%	37.36%
	0.5	Exponential	0.4049	0.3634	0.2533	0.2180	37.43%	40.01%
	0.5	Quadratic	0.3440	0.2856	0.2585	0.2094	24.84%	26.69%

Table 5: Comparison of the average lengths of the 95% credible intervals from the complete part of the data (D_1) and the 95% credible interval estimated after taking into account the missing part of the data(D_2). D stands for the whole dataset given by $D = D_1 \cup D_2$. The percentage improvement in the length of the interval is reported in last two columns. Here, 40% of observations have X_1 missing.

MAR covariates, and comparison of these two tables show that improvements in widths are more for Table 5 with a higher proportion of observations with missing covariates. It is important to note that even though our simulation model allows only the X_1 to be MAR, the lengths of the credible intervals for both the parameters θ_1 and θ_2 improve when using the missing data extension of BSIM. This may be due to the strong posterior dependence of the parameters θ_1 and θ_2. The improvements in the widths of interval estimates also increase with the increase in error standard deviation (σ).

In our simulation study 6, we compare our proposed BSIM extension method to MAR covariates with the inverse probability weighted outer product gradient (OPG) method (called NOPG in short), described in Guo et al. (2014) and based on the estimates $\hat{\alpha}$ of the index vector. Guo et al. (2014) suggested either a kernel-based, nonparametric method or logistic regression to find the probability weights. We use the nonparametric, kernel-based methods to compute the probability weights because the logistic regression-based approach run into numerical issues when the sample size n is 50 (small). We use the median of the sum of square error (MSSE) of $\hat{\alpha}$ to compare our Bayesian method with Guo et al. (2014)'s NOPG method. The results are tabulated in Table 6. Based on these results, our proposed Bayesian method either performs equally well or better when the error standard deviation σ is small. However, for higher values of σ, NOPG performs better than the proposed method in terms of MSSE of $\hat{\alpha}$. However, it is worth noting that computing the 95% confidence interval of α and $f(\alpha^{\mathrm{T}}X)$ are computationally difficult for the NOPG method in spite of the asymptotic results in Guo et al. (2014). In comparison, the extension of the Bayesian method to the MAR covariate provides the posterior credible intervals for the parameters of interest, making the Bayesian method more useful in practice.

			20% Missing		40% Missing	
n	σ	True f	BSIM	NOPG	BSIM	NOPG
50	0.01	Exponential	0.0003	0.0008	0.0006	0.0007
	0.01	Quadratic	0.0003	0.0002	0.0005	0.0002
	0.1	Exponential	0.0007	0.0006	0.0011	0.0006
	0.1	Quadratic	0.0005	0.0003	0.0009	0.0002
	0.5	Exponential	0.0109	0.0047	0.0171	0.0053
	0.5	Quadratic	0.0093	0.0033	0.0092	0.0048
100	0.01	Exponential	0.0002	0.0002	0.0006	0.0002
	0.01	Quadratic	0.0003	0.0001	0.0001	0.0001
	0.1	Exponential	0.00038	0.0002	0.0007	0.0002
	0.1	Quadratic	0.0006	0.0001	0.0022	0.0001
	0.5	Exponential	0.0086	0.0014	0.0083	0.0013
	0.5	Quadratic	0.0102	0.0016	0.0089	0.0011

Table 6: Comparison of the median of sum of square error (MSSE) of $\hat{\alpha}$ between the proposed method to deal with missing data and method described in Guo et al. (2014).

In simulation study 7, we investigate the 95% coverage probabilities using the complete data (D_1), and how it changes when including the data with missing covariates (D_2). We keep the same tuning parameters used in previous simulation scenarios. The

			Coverage Probabilities			
			Using D_1		Using $D = D_1 \cup D_2$	
n	σ	True f	θ_1	θ_2	θ_1	θ_2
50	0.01	Exponential	90%	92%	92%	92%
	0.01	Quadratic	100%	96%	100%	98%
	0.1	Exponential	86%	86%	88%	86%
	0.1	Quadratic	100%	96%	98%	94%
100	0.01	Exponential	100%	98%	100%	98%
	0.01	Quadratic	100%	100%	94%	86%
	0.1	Exponential	96%	100%	96%	96%
	0.1	Quadratic	100%	98%	96%	78%

Table 7: 95% Coverage probability of θ_1, θ_2 when only D_1 or both D_1, D_2 are used. In this simulation, 20% of the observations have X_1 missing.

			Coverage Probabilities			
			Using D_1		Using $D = D_1 \cup D_2$	
n	σ	True f	θ_1	θ_2	θ_1	θ_2
50	0.01	Exponential	86%	86%	86%	84%
	0.01	Quadratic	98%	94%	94%	78%
	0.1	Exponential	84%	80%	84%	78%
	0.1	Quadratic	90%	90%	90%	84%
100	0.01	Exponential	100%	96%	82%	56%
	0.01	Quadratic	100%	100%	44%	28%
	0.1	Exponential	92%	90%	90%	70%
	0.1	Quadratic	100%	100%	72%	42%

Table 8: 95% Coverage probability of θ_1, θ_2 when only D_1 or both D_1 and D_2 are used. In this simulation, 40% of the observations have X_1 missing.

results, when 20% of the observations have X_1 missing, are in Table 7. For 40% of observations having X_1 missing, the results are summarized in Table 8. We observe that when using $D = D_1 \cup D_2$, the coverage probability drops compared to when only D_1 is used for inference. Moreover, when the percentage of missing data is higher, the coverage probability decreases.

We also compare the predictive performance of the proposed method with other existing methods (CSOPG, KSIR, Antoniadis-SIM and Gramacy-SIM). These results are based on simulation study 8. Here, we fixed the sample size of the overall data and use 80% of the data for training the model. The remaining 20% of the data is used to evaluate the predictive performance of the fitted model. This method is replicated 50 times. The median of the mean absolute errors of the predictions $(\frac{1}{n_{\text{test}}}\sum_{i=1}^{n}|f_0(\alpha_0^{\mathrm{T}}x_i^{\text{test}}) - \hat{f}(\hat{\alpha}^{\mathrm{T}}x_i^{\text{test}}|)$ are reported in Table 9. Here, x_i^{test} denotes the covariate of the i^{th} subject in the test set. n_{test} denotes the number of observations in the test set. When using Bayesian methods, namely Antoniadis-SIM, Gramacy-SIM and BSIM, we use a posterior predictive median for each subject as a final prediction. We observe that the proposed method outperforms the existing methods when the error variance is small

n	σ	True f	CSOPG	KSIR	BSIM	Antoniadis-SIM	Gramacy-SIM
50	0.01	Exponential	0.2087	0.2175	0.0805	0.6255	0.2584
	0.01	Quadratic	0.3357	0.5879	0.0721	1.0112	0.1313
	0.1	Exponential	0.2423	0.2623	0.1491	1.0053	0.3543
	0.1	Quadratic	0.3409	0.6249	0.1552	0.9643	0.2671
	0.3	Exponential	0.3476	0.3718	0.3591	1.1358	0.3858
	0.3	Quadratic	0.4116	0.5909	0.3764	0.9607	0.3946
	0.5	Exponential	0.5476	0.5395	0.6615	0.8733	0.6335
	0.5	Quadratic	0.5510	0.7419	0.5966	1.1024	0.5966
100	0.01	Exponential	0.2382	0.2560	0.0905	0.4656	0.0913
	0.01	Quadratic	0.3590	0.4502	0.1021	0.7287	0.0939
	0.1	Exponential	0.2564	0.2592	0.1302	0.5034	0.1203
	0.1	Quadratic	0.3569	0.4346	0.1817	0.7238	0.1769
	0.3	Exponential	0.3608	0.3590	0.3850	0.5038	0.2206
	0.3	Quadratic	0.4148	0.6092	0.3625	0.8967	0.3638
	0.5	Exponential	0.5013	0.5080	0.6109	0.6637	0.4832
	0.5	Quadratic	0.5000	0.6223	0.5609	0.9140	0.5184

Table 9: Comparison of predictive performance of the existing methods (CSOPG, KSIR, Antoniadis et al. (2004), Gramacy and Lian (2012) and the proposed method (BSIM). We report the median of the mean absolute errors of prediction based on 50 replicates. 80% of the data was used to train the models and the rest 20% was used for evaluating prediction performance.

($\sigma = 0.01, 0.1, 0.3$). When the error variance is large, the proposed method has comparable performance to existing methods.

7 Analysis of Air Quality Data

We illustrate our method via reanalysis of the air quality study (Chambers, 1983), which was the motivating study for many major methodological developments in SIM literature (Hristache et al., 2001; Antoniadis et al., 2004; Choi et al., 2011). The full dataset, available via the `datasets` package in `R`, contains the logarithm of daily concentration (response) as well as three predictors (covariates) including the solar radiation, wind speed and temperature of New York City for 153 days from May to September 1973. Out of these 153 samples, there are 42 data points with at least one covariate or response missing. We first use the frequentist diagnostic described in Guo et al. (2016) to check whether a SIM is appropriate for modeling the response as a function of three covariates. The observed test-statistic of $T^{\text{DEE}} = 5.77$ results in a small p-value, indicating that it is reasonable to use a SIM. We also provide an exploratory assessment of the adequacy of our Bayesian SIM using the empirical cumulative distribution function (cdf) $C(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i \leq t)$ versus the fitted cdf $M(t) = \frac{1}{n}\sum_{i=1}^{n} \Phi((t - \hat{y}_i)/\hat{\sigma})$, where $\mathbb{I}(\cdot)$ denotes the indicator function, $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, $\hat{y}_i$ and $\hat{\sigma}$ denote the Bayesian estimates of $E[Y_i|X_i] = f(\alpha^{\mathrm{T}} X_i)$ and σ, respectively (computed via MCMC samples from the joint posterior). Validity of (2.1) is supported by the agreement between $C(t)$ and $M(t)$ across the range of observed

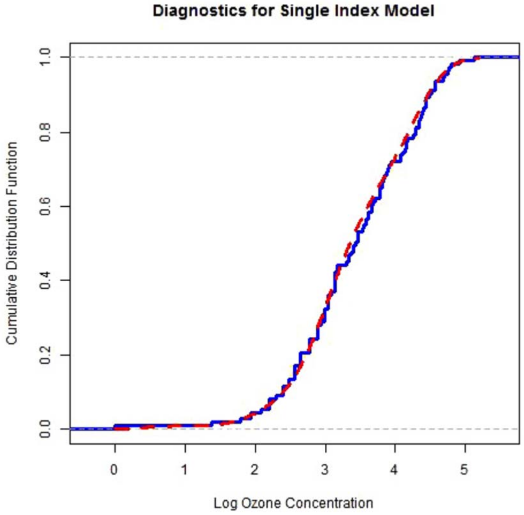

Figure 1: Model diagnostics to check whether a Bayesian SIM is appropriate for the air-quality study. The solid blue curve is the empirical cumulative distribution $C(t)$ of the observed response, and the red dashed line represents the fitted cumulative distribution function $M(t)$ under SIM. These two curves are almost indistinguishable, indicating the adequacy of Bayesian SIM for the air-quality study.

responses, t. In Figure 1, we plotted $C(t)$ as the solid blue curve and $M(t)$ as the red dashed curve. The plot reveals that these two curves are nearly identical over the range of all observed responses, indicating that a SIM is appropriate for this study across all observed response values.

Compared to previous methods that do not handle missing data, we first analyze using 111 data points used by previous methods. For our Bayesian method, the hyper-parameters of beta proposal density are $c_1 = c_2 = 5000$ for both coordinates of θ. The prior distribution on error variance σ^2 is inverse gamma with shape parameter 15 and rate parameter 0.06, corresponding to the prior guess of around 0.02 for σ (because the true σ is typically believed to be small for such a semiparametric regression model). The link function $f(\cdot)$ is assigned an OU process prior with the hyper-parameter κ having uniform hyper-prior on a support of grid of points in $[0.1, 3]$ with a common grid interval width 0.05. Similar to the simulation studies, we use 3,000 MCMC iterations (after discarding the first 2,000 iterations for burn-in) to compute the Monte Carlo approximation of the Bayesian estimates and posterior standard deviations of $(\alpha_1, \alpha_2, \alpha_3)$. Table 10 presents the point estimates of α, obtained from our method as well as three other competing methods.

As is apparent from Table 10, the values of point estimates of α from our Bayesian methods are close to the corresponding estimates obtained from existing methodologies. Unlike the method of Choi et al. (2011), we also provide the posterior standard deviations (a measure of uncertainty in Bayesian estimates) of $\alpha_1, \alpha_2, \alpha_3$, and these posteriors

Method	Point Estimate	Standard Deviation
Hristache et al. (2001)	(0.0407, 0.5263, -0.8493)	(0.0821, 0.1469, 0.0886)
Antoniadis et al. (2004)	(0.0817, 0.5565, -0.8103)	(0.0677, 0.1248, 0.0832)
Choi et al. (2011)	(0.0295, 0.5714, -0.8202)	-
BSIM with D_1	(0.0506, 0.4859, -0.8725)	(0.0006, 0.012, 0.006)
BSIM with $D = D_1 \cup D_2$	(0.0506, 0.4863, -0.8722)	(0.0006, 0.012, 0.006)

Table 10: Estimates of index vector α obtained from five competing analysis methods. The estimates reported here correspond to solar radiation, wind speed and temperature, respectively.

standard deviations are substantially smaller than the standard errors for competing frequentist estimates. This reanalysis demonstrates the feasibility and advantages of our method compared to existing SIM tools in terms of ease of computation, convergence rate and small autocorrelations among MCMC samples, as well as the ability to handle missing covariates.

In Figure 2, we overlay four estimates of the link function $f(z)$ versus index z obtained from three previous methods (Hristache et al., 2001; Antoniadis et al., 2004; Choi et al., 2011) and our Bayesian method. For three existing methods, only the estimates of α have been provided without any explicit estimate of $f(\cdot)$. To make all four methods comparable, we estimate the function $f(\cdot)$ for these three methods using smoothing

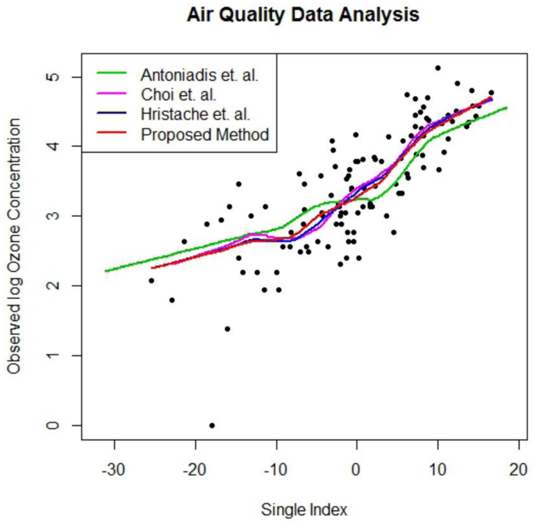

Figure 2: Plots of the estimated $f(z)$ versus index z obtained from our Bayesian and competing methods. The scatterplots are observed y_i versus estimated $\alpha^{\mathrm{T}} x_i$ from our Bayesian method. Estimated $f(\cdot)$ obtained from our Bayesian methods is similar to the Bayesian estimates from Antoniadis et al. (2004) and Choi et al. (2011), but all Bayesian estimates are different from the frequentist estimate from Hristache et al. (2001).

splines on the estimated $\alpha^{\mathrm{T}} x_i$. Estimated $f(\cdot)$ from our Bayesian method is very similar to the estimates from other two existing Bayesian methods of Choi et al. (2011) and Antoniadis et al. (2004). The frequentist estimate of $f(\cdot)$ (Hristache et al., 2001) appears to be somewhat different from the estimates from all three Bayesian methods.

As mentioned earlier, the air-quality study has 5 observations with covariate solar radiation missing. The point estimate of α obtained from our proposed missing covariate method (last line of Table 10) is very close to the point estimate obtained from analysis using only the completely observed data points. Since only 4.5% of the samples have missing covariates, the method accommodating the missing covariate mechanism did not result in a substantial change in the values of the Bayesian point estimates. Table 10 shows the posterior standard deviations of the index vector α with and without using the observations with missing covariates in the analysis (last two rows of Table 10). Since only a small part of the sample has missing covariates, the posterior standard deviations did not change substantially after incorporating the data points with missing covariates.

We also evaluate the performance of our proposed method for predictions. For this evaluation, we have ignored the 5 observations with missing covariates and used only the remaining 111 observations. We use a 5-fold, cross-validation of the data to provide the mean sum of squares of error for the proposed method and compare it with existing methods. Our Bayesian method has used the proposal density with the tuning parameters $c_1 = c_2 = 175$. The results are provided in Table 11. BSIM performs better than Choi-SIM and is comparable with Antoniadis-SIM. Gramacy-SIM has better prediction performance than BSIM. However, Gramacy-SIM does not estimate the index vector α as precisely as BSIM. Hence, BSIM provides a precise estimate of the index vector without compromising the predictive performance.

CSOPG	KSIR	HJS	Antoniadis-SIM	Gramacy-SIM	BSIM	Choi-SIM
0.3933	0.3896	0.3401	0.4587	0.3973	0.4539	0.6985

Table 11: Comparison of median of mean absolute error of prediction based on 5-fold cross-validation. We compare our proposed method (BSIM) with CSOPG, KSIR, Antoniadis et al. (2004), Choi et al. (2011), Gramacy and Lian (2012), Hristache et al. (2001). The table shows that the performance of the proposed method is comparable to the existing Bayesian methods.

8 Conclusion

It is straightforward to extend our method to other priors on $f(\cdot)$, including a Gaussian process with other types of covariance kernels. However, these extensions may come at an additional cost of computation owing to the lack of closed-form expressions of the inverse and determinant of the correlation matrix.

Even though our data example has only one covariate missing, it is straightforward to extend our approach to more than one and even discrete valued covariates subject to missing-at-random. Even though it is beyond the scope of this paper, our computational method can be extended to handle even nonignorable missing mechanisms. We found that via incorporating observation with missing covariates in data analysis, we can improve the precision/width of the Bayesian estimates, but with some potential bias trade-off.

For the first time in the Bayesian single index model framework, we provide theoretical guarantees in terms of posterior convergence rates of the overall regression function $g(x) = f(\alpha^{\mathrm{T}} x)$. The rate is optimal as long as the inverse bandwidth parameter is chosen appropriately. An immediate followup of our theoretical result is to show that the marginal posterior of the index vector α converges at a parametric rate.

We intend to extend our Bayesian methods to the partial linear SIM with regression function (conditional mean response) $E(Y \mid X, Z) = \alpha_1^{\mathrm{T}} Z + f(\alpha_2^{\mathrm{T}} X)$, where X and Z are two covariate vectors with Z having a linear effect and X having a nonlinear effect on the mean response. An important future research direction is toward modeling different quantiles of responses using Bayesian single index models. The fundamental challenge here is to obtain an appropriate stochastic model of the response that allows flexible skewness.

Supplementary Material

Supplementary material for "A New Bayesian Single Index Model with or without Covariates Missing at Random" (DOI: 10.1214/19-BA1170SUPPa; .pdf).

Supplementary material for "A New Bayesian Single Index Model with or without Covariates Missing at Random" (DOI: 10.1214/19-BA1170SUPPb; .zip).

References

Antoniadis, A., Grégoire, G., and McKeague, I. W. (2004). "Bayesian estimation in single-index models." *Statistica Sinica*, 14(4): 1147–1164. MR2126345. 760, 769, 774, 776, 777

Chambers, J. M. (1983). *Graphical methods for data analysis*. 759, 761, 774

Choi, T., Shi, J. Q., and Wang, B. (2011). "A Gaussian process regression approach to a single-index model." *Journal of Nonparametric Statistics*, 23(1): 21–36. MR2780813. doi: https://doi.org/10.1080/10485251003768019. 760, 765, 769, 774, 775, 776, 777

Dhara, K., Lipsitz, S., Pati, D., and Sinha, D. (2019). "Supplementary materials for "A New Bayesian Single Index Model with or without Covariates Missing at Random"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1170SUPPa. doi: https://doi.org/10.1214/19-BA1170SUPPb. 766, 769

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin. MR1851606. doi: https://doi.org/10.1007/978-0-387-21606-5. 762

Gramacy, R. B. and Lian, H. (2012). "Gaussian process single-index models as emulators for computer experiments." *Technometrics*, 54(1): 30–41. MR2904733. doi: https://doi.org/10.1080/00401706.2012.650527. 760, 769, 774, 777

Guo, X., Niu, C., Yang, Y., and Xu, W. (2015). "Empirical likelihood for single index model with missing covariates at random." *Statistics*, 49(3): 588–601. MR3349080. doi: https://doi.org/10.1080/02331888.2014.881826. 761

Guo, X., Wang, T., and Zhu, L. (2016). "Model checking for parametric single-index models: a dimension reduction model-adaptive approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 1013–1035. MR3557187. doi: https://doi.org/10.1111/rssb.12147. 774

Guo, X., Xu, W., and Zhu, L. (2014). "Multi-index regression models with missing covariates at random." *Journal of Multivariate Analysis*, 123: 345–363. MR3130439. doi: https://doi.org/10.1016/j.jmva.2013.10.006. 772

Hardle, W., Hall, P., Ichimura, H., et al. (1993). "Optimal smoothing in single-index models." *The Annals of Statistics*, 21(1): 157–178. MR1212171. doi: https://doi.org/10.1214/aos/1176349020. 760

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). "Direct estimation of the index coefficient in a single-index model." *The Annals of Statistics*, 595–623. MR1865333. doi: https://doi.org/10.1214/aos/1009210681. 760, 774, 776, 777

Ichimura, H. (1993). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models." *Journal of Econometrics*, 58(1): 71–120. MR1230981. doi: https://doi.org/10.1016/0304-4076(93)90114-K. 760

Li, K.-C. (1991). "Sliced inverse regression for dimension reduction." *Journal of the American Statistical Association*, 86(414): 316–327. MR1137117. 760, 769

Li, K.-C. and Duan, N. (1989). "Regression analysis under link violation." *The Annals of Statistics*, 1009–1052. MR1015136. doi: https://doi.org/10.1214/aos/1176347254. 760

Li, T. and Yang, H. (2016). "Inverse probability weighted estimators for single-index models with missing covariates." *Communications in Statistics-Theory and Methods*, 45(5): 1199–1214. MR3462142. doi: https://doi.org/10.1080/03610926.2012.705208. 761

Lin, W. and Kulasekera, K. (2007). "Identifiability of single-index models and additive-index models." *Biometrika*, 94(2): 496–501. MR2380574. doi: https://doi.org/10.1093/biomet/asm029. 762

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons. MR1925014. doi: https://doi.org/10.1002/9781119013563. 766, 767

Murray, I. and Adams, R. P. (2010). "Slice sampling covariance hyperparameters of latent Gaussian models." In *Advances in Neural Information Processing Systems*, 1732–1740. 764

Niu, C., Zhu, L., et al. (2017). "An adaptive-to-model test for parametric single-index models with missing responses." *Electronic Journal of Statistics*, 11(1): 1491–1526. MR3635920. doi: https://doi.org/10.1214/17-EJS1257. 761

Park, C. G., Vannucci, M., and Hart, J. D. (2005). "Bayesian methods for wavelet series in single-index models." *Journal of Computational and Graphical Statistics*, 14(4). MR2211366. doi: https://doi.org/10.1198/106186005X79007. 760, 762

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). "Semiparametric estimation of index coefficients." *Econometrica: Journal of the Econometric Society*, 1403–1430. MR1035117. doi: https://doi.org/10.2307/1913713. 760

Stoker, T. M. (1986). "Consistent estimation of scaled coefficients." *Econometrica: Journal of the Econometric Society*, 1461–1481. MR0868152. doi: https://doi.org/10.2307/1914309. 759, 760

Wang, H.-B. (2009). "Bayesian estimation and variable selection for single index models." *Computational Statistics & Data Analysis*, 53(7): 2617–2627. MR2665912. doi: https://doi.org/10.1016/j.csda.2008.12.010. 760

Xia, Y. (2006). "Asymptotic distributions for two estimators of the single-index model." *Econometric Theory*, 22(6): 1112–1137. MR2328530. doi: https://doi.org/10.1017/S0266466606060531. 760, 769

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002). "An adaptive estimation of dimension reduction space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 363–410. MR1924297. doi: https://doi.org/10.1111/1467-9868.03411. 760

Xue, L. (2013). "Estimation and empirical likelihood for single-index models with missing data in the covariates." *Computational Statistics & Data Analysis*, 68: 82–97. MR3103764. doi: https://doi.org/10.1016/j.csda.2013.06.017. 761

Acknowledgments

The authors would like to acknowledge the editor and the referees for their constructive suggestions to improve this paper. The authors would also like to thank Drs. Ian McKeague, Xu Guo, Lixing Zhu, Taeryon Choi and Robert Gramacy for providing the codes for their methods, which facilitated the comparisons with our method.

Dr. Pati acknowledges support for this project from the National Science Foundation (NSF DMS 1613156 & NSF DMS 1854731).