Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Bayesian Scalar on Image Regression With Nonignorable Nonresponse

Xiangnan Feng , Tengfei Li , Xinyuan Song & Hongtu Zhu

To cite this article: Xiangnan Feng , Tengfei Li , Xinyuan Song & Hongtu Zhu (2020) Bayesian Scalar on Image Regression With Nonignorable Nonresponse, Journal of the American Statistical Association, 115:532, 1574-1597, DOI: 10.1080/01621459.2019.1686391

To link to this article: https://doi.org/10.1080/01621459.2019.1686391

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 1574–1597: Applications and Case Studies
https://doi.org/10.1080/01621459.2019.1686391

Bayesian Scalar on Image Regression With Nonignorable Nonresponse

Xiangnan Feng*[a], Tengfei Li*[b], Xinyuan Song[c], and Hongtu Zhu[d]

[a]School of Economics and Management, Southwest Jiaotong University, Chengdu, China; [b]Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC; [c]Department of Statistics, Chinese University of Hong Kong, Shatin, NT, Hong Kong; [d]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC

ABSTRACT

Medical imaging has become an increasingly important tool in screening, diagnosis, prognosis, and treatment of various diseases given its information visualization and quantitative assessment. The aim of this article is to develop a Bayesian scalar-on-image regression model to integrate high-dimensional imaging data and clinical data to predict cognitive, behavioral, or emotional outcomes, while allowing for nonignorable missing outcomes. Such a nonignorable nonresponse consideration is motivated by examining the association between baseline characteristics and cognitive abilities for 802 Alzheimer patients enrolled in the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI1), for which data are partially missing. Ignoring such missing data may distort the accuracy of statistical inference and provoke misleading results. To address this issue, we propose an imaging exponential tilting model to delineate the data missing mechanism and incorporate an instrumental variable to facilitate model identifiability followed by a Bayesian framework with Markov chain Monte Carlo algorithms to conduct statistical inference. This approach is validated in simulation studies where both the finite sample performance and asymptotic properties are evaluated and compared with the model with fully observed data and that with a misspecified ignorable missing mechanism. Our proposed methods are finally carried out on the ADNI1 dataset, which turns out to capture both of those clinical risk factors and imaging regions consistent with the existing literature that exhibits clinical significance. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received December 2017
Accepted October 2019

KEYWORDS
Bayesian approach; Imaging data; Instrumental variable; Markov chain Monte Carlo; Nonignorable nonresponse

1. Introduction

The present study is motivated by a dataset extracted from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Since its launch in 2004, ADNI collected imaging, clinical, and laboratory data at multiple time points from cognitively normal controls (CN) and subjects with mild cognitive impairment (MCI) or Alzheimer's disease (AD). ADNI initially recruited approximately 800 subjects (ADNI-1) according to its initial aims and was extended by three follow-up studies, namely, ADNI-GO, ADNI-2, and ADNI-3. The overall goal of ADNI is to discover, optimize, standardize, and validate clinical trial measures and biomarkers used in AD research by determining the relationships between the clinical, cognitive, imaging, genetic, and biochemical biomarker characteristics for the entire spectrum of AD. More information on ADNI can be obtained at the official website (*www.adni-info.org*).

The primary objective of this study is to examine whether patients' numerous baseline biomarkers (e.g., structural imaging) can accurately predict their cognitive decline. The ability to accurately predict the rate of cognitive decline is critical for effective trial design for developing therapies for AD prevention and treatment, but the utility of these baseline biomarkers for such accurate prediction is not well established (Allen et al. 2016; Weiner et al. 2017a, 2017b). We consider the learning score of the Rey auditory verbal learning test (RAVLT), which is a widely used neuropsychological evaluation method that tests episodic declarative memory. The RAVLT learning scores of each subject were obtained at baseline and every 6 months thereafter across multiple study phases. We consider the RAVLT scores measured at the 36th month as the primary clinical outcome and the demographic, imaging, and clinical variables measured at baseline as predictors. To build an accurate predictive model for the RAVLT score, we have to appropriately deal with at least two challenging issues, including (I) missing RAVLT scores, particularly when the missing mechanism is nonignorable, and (II) high-dimensional imaging data.

The first challenge is that the nonresponse rate of the RAVLT score increases over time and attains a level of 45.6% at the 36th month, while the missing mechanism is nonignorable. To justify the missingness assumption, we divide the samples into two groups according to whether they are nonrespondents (Group 1) or respondents (Group 2) at the 36th month. Figure 1 summarizes the learning scores of both groups at baseline. Apparently, the subjects in Group 1 have considerably lower learning ability than those in Group 2. This finding implies that low learning ability at baseline is negatively associated with nonresponse probability at the 36th month. That is, elderly adults with weak cognitive ability tend to drop out early from the follow-up study.

CONTACT Hongtu Zhu *htzhu@email.unc.edu* Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516.
*Drs. Feng and Li are the joint first author of this article.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
These materials were reviewed for reproducibility.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

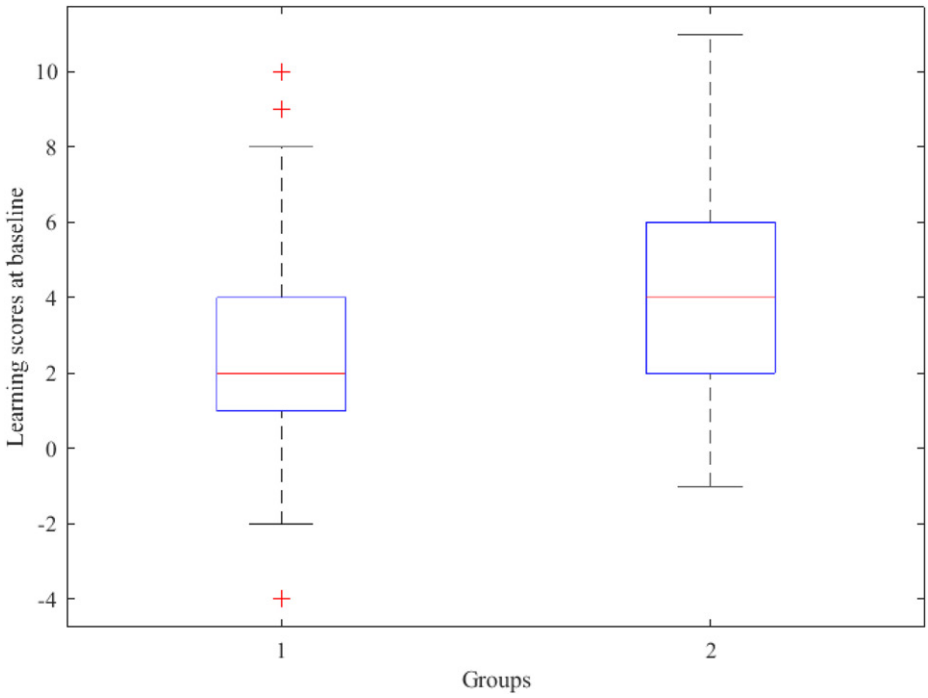

Figure 1. Learning scores at the baseline of the subjects from two groups in the ADNI dataset, where Group 1 includes the patients who exhibit missing learning scores at the 36th month and Group 2 contains the patients with observed learning scores at the same month.

Thus, the missing data mechanism is likely to be nonignorable and a missing data model should be considered to identify possible effects of learning ability together with other imaging and scalar covariates on the probability of data missingness.

In large-scale longitudinal neuroimaging studies, follow-up clinical outcomes are frequently missing from the dataset. Thus, appropriately managing nonresponse is of great importance. A nonresponse is regarded as ignorable when its probability is independent of the missing values (Little and Rubin 2002). However, the probability of nonresponse often depends on the observed and missing observations, and disregarding such a missing mechanism may destroy the representativeness of the remaining samples and subsequently lead to biased estimation results (Baker and Laird 1988; Diggle and Kenward 1994; Ibrahim, Lipsitz, and Chen 1999; Ibrahim, Chen, and Lipsitz 2001; Molenberghs and Kenward 2007). Modeling nonignorable missingness is challenging because the missing mechanism is generally unknown and may elicit additional model identifiability issues (Chen 2001; Qin, Leung, and Shao 2002; Tang, Little, and Raghunathan 2003; Ibrahim et al. 2005). Recently, two advanced methods have been proposed to facilitate model identification when dealing with nonignorable missingness under the exponential tilting model proposed by Kim and Yu (2011). The first method, which was developed by Kim and Yu (2011), relies on a set of external data obtained from an independent study, where further responses can be obtained in a subset of nonrespondents (see also Zhao, Zhao, and Tang 2013; Tang, Zhao, and Zhu 2014). Through the use of external data, the tilting parameter can be estimated, and the exponential tilting model for the formulation of nonresponse is well identified. However, such an external dataset is often unavailable in practice, making the procedure infeasible. The second method can address such a problem by introducing an instrumental variable, such as a covariate associated with the response but conditionally independent of the probability of data missingness. The advantage of this method has been demonstrated in recent works (Wang, Shao, and Kim 2014; Yang et al. 2014; Zhao and Shao 2015; Shao and Wang 2016). However, these methods are limited to modeling scalar responses and predictors.

The second challenge is the use of high-dimensional medical images (or functional data) observed at a set of grid points to accurately predict clinical outcomes. Many imaging studies have collected high-dimensional imaging data, such as magnetic resonance imaging (MRI) data and computed tomography, to extract useful information associated with the pathophysiology of various diseases, such as lung cancer and AD. Such information can further facilitate clinical decision-making (Gillies, Kinahan, and Hricak 2015). For instance, data obtained from MRI-based investigations may greatly contribute to the discovery and validation of prognostic biomarkers used to identify subjects at great risk of cognitive decline, thereby aiding researchers and clinicians in monitoring the progression of MCI and early AD, as well as developing new treatments and reducing the time and cost of clinical trials.

A functional linear model (FLM) and its variations have received extensive attention in the last two decades as popular predictive models based on functional predictors (Ramsay and Silverman 2005; Ferraty and Vieu 2006; Horváth and Kokoszka 2012; Morris 2015; Fraiman, Gimenez, and Svarc 2016; Wang, Chiou, and Mueller 2016). Many estimation methods have been developed to estimate the coefficient function of the FLM and its variations, but they differ in terms of the choice of basis or some combination thereof, and the approach to regularization (Morris 2015; Wang, Chiou, and Mueller 2016; Wang and Zhu 2017). The most common choices for basis include functional principal components, splines, and wavelets, among many others (Ramsay and Silverman 2005; Ferraty and Vieu 2006; Hall and Horowitz 2007; Yuan and Cai 2010). Functional principal

component analysis (FPCA) is an important tool that reduces the dimensionality of functional data (Müller and Stadtmüller 2005; Yao, Müller, and Wang 2005; Reiss and Ogden 2007, 2010; Goldsmith et al. 2011, among many others). Furthermore, FLM analysis is feasible given the connection between FPCA and ordinary linear mixed models (James 2002).

Although missing data problems have been extensively investigated, minimal work has focused on the analysis of functional data with missing scalar clinical outcomes. Preda, Saporta, and Mbarek (2010) considered a nonlinear iterative partial least squares method to accommodate functional predictors subject to data missingness. Gertheiss et al. (2013) conducted longitudinal scalar-on-function regression, which allows for ignorable missingness in functional regressors. Ferraty, Sued, and Vieu (2013) studied the mean estimation problems for scalar-on-function regression with ignorable nonresponse. Their study was extended by Ling, Liang, and Vieu (2015), who considered stationary ergodic functional processes as predictors with ignorable nonresponse. Chiou et al. (2014) modeled traffic monitoring data as functional processes and imputed missing values in the functional data by using a conditional expectation approach. However, the aforementioned developments focused only on functional data with ignorable missingness, and none of them considered ultrahigh-dimensional imaging data in the presence of nonignorable missingness. To the best of our knowledge, the study by Li et al. (2018) is the only article that developed a functional linear model for the joint modeling of functional predictors and nonignorable missing clinical outcomes. The methodology they proposed is a frequentist method that depends on an external dataset, which is rarely available in real applications.

The aim of this article is to develop a Bayesian scalar-on-image (BSOI) regression model that uses high-dimensional imaging data and other scalar variables as explanatory covariates to predict clinical outcomes that are not fully observed. The proposed approach comprises two stages. The first stage uses FPCA to extract the principal directions of variation in large-scale neuroimaging data, and the extraction is performed through the singular value decomposition (SVD) technique (Zipunnikov et al. 2011a; Zhu, Fan, and Kong 2014; Wang, Chiou, and Mueller 2016). For simplicity, we use FPCA in the development of BSOI although it is easy to use some fixed basis functions, such as a B-spline and wavelet. The second stage incorporates the extracted major principal scores into the regression. Regarding the modeling of nonignorable missingness, we propose an imaging exponential tilting model that is analyzed jointly with the BSOI regression. The imaging predictors involved in the exponential tilting model can be similarly assessed through FPCA. An instrumental variable is introduced to facilitate the identification of the proposed model. We conduct a full Bayesian analysis, not only given its power and efficiency in managing complex models and data structures but also because it incorporates useful prior information. Appropriate prior distributions can add valuable information for the inference of the missing mechanism and thus assist model identification and estimation (Ibrahim, Chen, and Lipsitz 2002). For instance, if a set of external data is available for preliminary analysis on the missing data model, the estimation results can be incorporated as prior inputs into the Bayesian analysis to improve the estimation accuracy.

The remainder of this article is organized as follows: Section 2 introduces the FPCA technique, a BSOI regression model, and an exponential tilting model with imaging predictors. Section 3 discusses how an instrumental variable improves model identification and estimation and develops a Bayesian approach with Markov chain Monte Carlo (MCMC) algorithms for statistical inference. Section 4 conducts simulations to examine the finite sample performance of the proposed method. Section 5 presents a comprehensive data analysis of the ADNI dataset presented above. Section 6 concludes the article with some discussions.

2. Model Description

2.1. FPCA for High-Dimensional Imaging Data

Suppose we have samples of imaging data $W_i(v)$ observed at V grid points v in a compact space $\mathcal{V}$ for $i = 1, \ldots, n$. The observed images are formulated through a functional model as follows:

$$W_i(v) = \mu(v) + X_i(v), \quad v \in \mathcal{V}, \tag{1}$$

where $\mu(v)$ is the mean image, and $X_i(v)$ is a centered second-order stochastic process with covariance operator $K_X(v_1, v_2) = E\{X_i(v_1)X_i(v_2)\}$. Similar to those of Zipunnikov et al. (2011a), the measurement errors of the imaging observations are not included in the model. This assumption is reasonable because the images are usually presmoothed. The Karhunen–Loeve expansion of the random process is based on the eigen-decomposition of the covariance operator, which yields

$$X_i(v) = \sum_{k=1}^{\infty} \xi_{ik}\psi_k(v), \tag{2}$$

where $\psi_k(v)$ represents the eigenimages of $K_X(v_1, v_2)$ for the multidimensional imaging data, ξ_{ik} denotes the uncorrelated eigenscores of the ith subject with nonincreasing variances λ_k, and λ_k is the eigenvalues of $K_X(v_1, v_2)$. FPCA commonly retains the first K eigenimages that account for most of the functional variability for the expansion (2) (Di et al. 2009; Zipunnikov et al. 2011a; Zhu, Fan, and Kong 2014; Wang, Chiou, and Mueller 2016). In practical applications, retaining a large number of eigenimages may lead to overfitting. Thus, we use the Bayesian information criterion (BIC) to select K for the analysis of the ADNI dataset in Section 5. The mean image $\mu(v)$ can be regarded as a zero vector by centralizing $W_i(v)$. Thus, the functional model (1) can be rewritten as follows:

$$W_i(v) = X_i(v) = \sum_{k=1}^{K} \xi_{ik}\psi_k(v). \tag{3}$$

Owing to the ultrahigh-dimensionality, eigen-decomposing the covariance operator of the functional process $X_i(v)$ is challenging. For example, we consider a two-dimensional image $\{X_i(v)\}$ with 300 grids on each dimension, that is, $V = 90{,}000$. This results in a covariance operator $\mathbf{K}_X = \{K_X(v_1, v_2)\}$ of dimension $90{,}000 \times 90{,}000$. Consequently, a brutal-force eigenanalysis on $\mathbf{K}_X$ requires $O(V^3)$ operations, which are essentially impossible. To address such a problem, Zipunnikov et al. (2011a) developed an FPCA procedure based on SVD for high-dimensional data. The method exploits the relationship between

the SVD on $X_i(v)$ and the eigen-decomposition on $\mathbf{K}_X$. It also exploits the advantage that the number of subjects, n, is usually small to modest. Specifically, we consider a $V \times n$ matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, whose rank is at most $\min(n, V)$, $\mathbf{X}_i = (X_i(v) : \text{all grid points } v \in \mathcal{V})$ is a $V \times 1$ vector for each i. The SVD of $\mathbf{X}$ can then be represented as

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \tag{4}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $V \times n$ and $n \times n$ unitary orthogonal matrices, respectively, and $\mathbf{S}$ is an $n \times n$ diagonal matrix with nonnegative singular values of $\mathbf{X}$ as its diagonal elements. The computational cost to obtain the SVD is $O(Vn^2 + n^3)$ (Golub and Loan 1996), which is much smaller than that required for the direct eigen-decomposition of the covariance operator. The eigenimage (eigenfunction) $\psi_k(v)$, eigenvalue λ_k of $\mathbf{K}_X$ and the eigenscores ξ_{ik} of the subjects can be calculated as follows. The eigenimage $\mathbf{\Psi}_k = \psi_k(v)$ is given by the kth column of $\mathbf{U}$. The eigenvalue λ_k equals s_k^2, where s_k is the kth diagonal element of $\mathbf{S}$. The eigenscores ξ_{ik} are given by the columns of $\mathbf{S}\mathbf{V}^T$ truncated to the first K coordinates. To implement the SVD, we first express the spectral decomposition of $\mathbf{X}^T\mathbf{X}$ as $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T$. Then, $\mathbf{U}$ can be calculated as $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{S}^{-1}$.

Notably, for an ultrahigh-dimensional $\mathbf{V}$, the centered imaging data matrix $\mathbf{X}$ may be extremely large and thus cannot be loaded into the computer memory. In such an instance, $\mathbf{X}$ can be partitioned into several blocks as $\mathbf{X}^T = [(\mathbf{X}^{(1)})^T|(\mathbf{X}^{(2)})^T|\ldots|(\mathbf{X}^{(M)})^T]$, where the number of blocks, M, can be selected to make each block adapt to the available computer memory so that it is feasible to calculate $\mathbf{X}^T\mathbf{X}$ and $\mathbf{U}$ as follows: $\mathbf{X}^T\mathbf{X} = \sum_{m=1}^{M}(\mathbf{X}^{(m)})^T\mathbf{X}^{(m)}$, $\mathbf{U}^{(m)} = \mathbf{X}^{(m)}\mathbf{V}\mathbf{S}^{-1}$, and $\mathbf{U}^T = [(\mathbf{U}^{(1)})^T|(\mathbf{U}^{(2)})^T|\ldots|(\mathbf{U}^{(M)})^T]$. Alternatively, we may employ the efficient algorithm of multidimensional FPCA from Chen and Jiang (2017).

2.2. BSOI Regression

We consider observations $\{y_i, \mathbf{z}_i, (X_i(v) : v \in \mathcal{V})\}$ for $i = 1, \ldots, n$ from n independent subjects, where y_i is the clinical variable of interest subject to missingness, $\mathbf{z}_i$ is a $Q \times 1$ vector of observed scalar covariates, and $X_i(v)$ represents the imaging data described above. The BSOI regression model is thus defined as follows:

$$y_i = \alpha + \int_{\mathcal{V}} X_i(v)\beta(v)dv + \boldsymbol{\gamma}^T\mathbf{z}_i + \delta_i, \tag{5}$$

where α is the intercept, $\beta(\cdot)$ is a coefficient image, $\boldsymbol{\gamma}$ is a $Q \times 1$ vector of the coefficients for the covariates, and δ_i is normal random noise with the variance parameter σ^2. The $\beta(v)$ is the coefficient corresponding to the vth voxel of the image. Thus, a natural interpretation of $\beta(\cdot)$ is that the regions of imaging data with large $|\beta(v)|$ have strong effects on the clinical outcome of interest.

With the eigenimages derived in the previous section, both $X_i(\cdot)$ and $\beta(\cdot)$ can be approximated by truncated Karhumen–Loeve expansions as $X_i(v) \approx \sum_{k=1}^{K}\xi_{ik}\psi_k(v)$ and $\beta(v) \approx \sum_{k=1}^{K}\beta_k\psi_k(v)$, where the β_ks' are the eigenbasis coefficients of $\beta(\cdot)$. Subsequently, model (5) can be rewritten as

$$y_i = \alpha + \sum_{k=1}^{K}\beta_k\xi_{ik} + \boldsymbol{\gamma}^T\mathbf{z}_i + \delta_i. \tag{6}$$

Therefore, the BSOI regression becomes a high-dimensional linear regression model, which can be readily analyzed.

2.3. Exponential Tilting Model for Nonignorable Missingness

We introduce an indicator variable r_i to model the missingness of y_i such that $r_i = 1$ if y_i is missing and $r_i = 0$ otherwise. Naturally, we assume a Bernoulli distribution of r_i as follows:

$$r_i|(y_i, X_i(\cdot), \mathbf{z}_i) \sim \text{Bernoulli}(\pi_i), \tag{7}$$

where $\pi_i = \pi(y_i, X_i(\cdot), \mathbf{z}_i)$ is the probability of missingness for y_i, and r_i and r_j are assumed to be independent for $i \neq j$. Furthermore, an exponential tilting model with imaging predictors is proposed for π_i as follows:

$$\begin{aligned}\pi_i &\equiv \Pr(r_i = 1|y_i, X_i(\cdot), \mathbf{z}_i) \\ &= \frac{\exp(\int_{\mathcal{V}} X_i(v)\beta_r(v)dv + \boldsymbol{\gamma}_r^T\mathbf{z}_i + \phi y_i)}{1 + \exp(\int_{\mathcal{V}} X_i(v)\beta_r(v)dv + \boldsymbol{\gamma}_r^T\mathbf{z}_i + \phi y_i)},\end{aligned} \tag{8}$$

where ϕ is the tilting parameter that determines the amount of departure from the ignorability of the missing mechanism. Model (8) can be regarded as an extension of the linear missing data model proposed by Ibrahim, Chen, and Lipsitz (2001). Furthermore, by using $\sum_{k=1}^{K}\beta_{rk}\psi_k(v)$ to approximate $\beta_r(v)$, model (8) reduces to

$$\begin{aligned}\pi_i &\equiv Pr(r_i = 1|y_i, X_i(v), \mathbf{z}_i) \\ &= \frac{\exp(\alpha_r + \sum_{k=1}^{K}\beta_{rk}\xi_{ik} + \boldsymbol{\gamma}_r^T\mathbf{z}_i + \phi y_i)}{1 + \exp(\alpha_r + \sum_{k=1}^{K}\beta_{rk}\xi_{ik} + \boldsymbol{\gamma}_r^T\mathbf{z}_i + \phi y_i)}.\end{aligned} \tag{9}$$

Finally, our BSOI model consists of Equations (6), (7), and (9).

3. Estimation

3.1. Identifiability

The identifiability of a nonignorable missing mechanism is often a challenging issue and thus requires careful investigation. As remarked by Lindley (1972), it is always possible to conduct a Bayesian analysis by assigning proper priors on model parameters regardless of the model identifiability. However, a practical consequence of a nonidentifiable model is that it may trap Bayesian implementations into the possibility of drifting to extreme values even with proper priors (Gelfand and Smith 1990). Therefore, achieving model identification is crucial for Bayesian analyses. The formal notion of Bayesian identifiability from Dawid (1979) has established that Bayesian nonidentifiability is equivalent to a lack of identifiability in the likelihood (Gelfand and Sahu 1999). This equivalence implies that the Bayesian model is identifiable as long as two different populations do not exhibit the same observed data likelihood. We thus investigate the identifiability of the model likelihood below.

The observed data likelihood function of our BSOI model is

$$\begin{aligned}&\prod_i [\Pr(r_i = 0|y_i, X_i(\cdot), \mathbf{z}_i)p(y_i|X_i(\cdot), \mathbf{z}_i)]^{1-r_i} \\ &\Big[\int \Pr(r_i = 1|y, X_i(\cdot), \mathbf{z}_i)p(y|X_i(\cdot), \mathbf{z}_i)dy\Big]^{r_i},\end{aligned} \tag{10}$$

where $\int \Pr(r_i = 1|y, X_i(\cdot), \mathbf{z}_i)p(y|X_i(\cdot), \mathbf{z}_i)dy$ is given by

$$\int \{1 - \Pr(r_i = 0|y, X_i(\cdot), \mathbf{z}_i)\}p(y|X_i(\cdot), \mathbf{z}_i)dy$$
$$= 1 - \int \Pr(r_i = 0|y, X_i(\cdot), \mathbf{z}_i)p(y|X_i(\cdot), \mathbf{z}_i)dy.$$

Thus, the joint model is identifiable when two different populations do not provide the same $\Pr(r = 0|y, X(\cdot), \mathbf{z})p(y|X(\cdot), \mathbf{z})$ for all possible $(y, X(\cdot), \mathbf{z})$. Although the proposed model assumes a parametric framework, identifiability remains nontrivial, as demonstrated by the following example.

Example 1. Suppose that models (5) and (9) exclude the functional covariate, while model (9) satisfies $\phi \neq 0$. Then, we have

$$\Pr(r = 0|y, z)p(y|z) = \frac{\exp\{-(y - \alpha - \gamma z)^2/2\sigma^2\}}{\sqrt{2\pi\sigma^2}\{1 + \exp(\alpha_r + \gamma_r z + \phi y)\}}.$$

Letting $\{\alpha, \gamma, \sigma, \alpha_r, \gamma_r, \phi\}$ and $\{\alpha', \gamma', \sigma', \alpha_r', \gamma_r', \phi'\}$ denote two sets of parameters, and

$$\frac{\exp\{-(y - \alpha - \gamma z)^2/2\sigma^2\}}{\sqrt{2\pi\sigma^2}\{1 + \exp(\alpha_r + \gamma_r z + \phi y)\}} = \frac{\exp\{-(y - \alpha' - \gamma' z)^2/2\sigma'^2\}}{\sqrt{2\pi\sigma'^2}\{1 + \exp(\alpha_r' + \gamma_r' z + \phi' y)\}}, \quad \forall (y, z), \tag{11}$$

it can be shown that model (11) holds if $\sigma' = \sigma$, $\alpha_r' = -\alpha_r$, $\gamma_r' = -\gamma_r$, $\phi' = -\phi$, $\alpha_r = \phi^2\sigma^2/2 - \phi\alpha$, $\alpha' = \alpha - \phi\sigma^2$, $\gamma' = \gamma$, $\phi = -\gamma_r/\gamma$, implying that the model is unidentifiable.

This simple example sheds new insights on the construction of identifiability conditions. Specifically, two possible solutions of (11) are given by

$$\text{Scenario } (i) \begin{cases} \exp\{-(y - \alpha - \gamma z)^2/2\sigma^2\}\sigma' \\ \quad = \exp\{-(y - \alpha' - \gamma' z)^2/2\sigma'^2\}\sigma, \\ \exp\{-(y - \alpha - \gamma z)^2/2\sigma^2\}\sigma' \\ \exp(\alpha_r' + \gamma_r' z + \phi' y) \\ \quad = \exp\{-(y - \alpha' - \gamma' z)^2/2\sigma'^2\}\sigma \\ \exp(\alpha_r + \gamma_r z + \phi y), \end{cases}$$

$$\text{Scenario } (ii) \begin{cases} \exp\{(y - \alpha - \gamma z)^2/2\sigma^2\}\sigma' \\ = \exp\{-(y - \alpha' - \gamma' z)^2/2\sigma'^2\}\sigma \\ \exp(\alpha_r + \gamma_r z + \phi y), \\ \exp\{-(y - \alpha' - \gamma' z)^2/2\sigma'^2\}\sigma \\ = \exp\{-(y - \alpha - \gamma z)^2/2\sigma^2\}\sigma' \\ \exp(\alpha_r' + \gamma_r' z + \phi' y). \end{cases}$$

If Scenario (i) holds for all (y, z) in their domain, then $(\alpha, \gamma, \sigma, \alpha_r, \gamma_r, \phi) = (\alpha', \gamma', \sigma', \alpha_r', \gamma_r', \phi')$ holds naturally. Then, the model is identifiable if one can exclude Scenario (ii). This can be achieved by introducing some additional assumptions. One such assumption is that a covariate u exists in $\mathbf{z} = (\mathbf{z}^{*T}, u)^T$ such that $\Pr(y|\mathbf{z})$ depends on u, whereas $\Pr(r = 0|y, \mathbf{z})$ does not. The covariate u is called a nonresponse instrument, which has been demonstrated to be essential for the identification and estimation of a nonignorable missing mechanism in regression without functional covariates (Wang, Shao, and Kim 2014; Shao and Wang 2016). As an illustration, we set $\mathbf{z}^*$ to be a nonzero constant and $u = z$ as the instrumental variable. In this case, the use of the instrumental variable $u = z$ implies $\gamma_r = 0$, yielding that $\exp(\alpha_r + \gamma_r z + \phi y)$ becomes $\exp(\alpha_r + \phi y)$ and is independent of z. However, given that $\exp\{-(y - \alpha - \gamma z)^2/2\sigma^2\}\sigma'/\exp\{-(y - \alpha' - \gamma' z)^2/2\sigma'^2\}/\sigma$ depends on z for any $(\alpha, \gamma, \sigma) \neq (\alpha', \gamma', \sigma')$, Scenario (ii) can be excluded and the model is identifiable.

In many real applications, it is not difficult to identify the instrumental variable. For example, we are interested in predicting monthly income, which is often only partially observed to protect privacy. It may be reasonable to assume that the probability of data missingness is independent of gender, age group, and educational level conditional on monthly income (Wang, Shao, and Kim 2014; Shao and Wang 2016). Since monthly income is usually associated with gender, age group, and educational level, we may choose any of the three covariates as the instrumental variable. Based on the existing literature (Wang, Shao, and Kim 2014; Zhao and Shao 2015; Shao and Wang 2016) and our experience, we have the following recommendations for choosing the instrumental variable. (i) The instrumental variable u has to be related to the response y and conditionally independent of the nonresponse probability. That is, $\Pr(y|X(\cdot), \mathbf{z}^*, u)$ should depend on u, whereas $\Pr(r = 0|y, X(\cdot), \mathbf{z}^*, u)$ should not. Experimental analyses that include all covariates in both the scalar on image regression and the exponential tilting model could provide clues on the covariates that have potential to serve as the instrumental variable. (ii) A sensitivity analysis with different choices of the instrumental variable may be considered.

We give a theoretical justification below in a more general setting and provide the proof in Appendix A. Specifically, we assume $\Pr(y|X(\cdot), \mathbf{z}) = \Pr(y|X(\cdot), \mathbf{z}; \boldsymbol{\theta}_y)$ and $\Pr(r = 0|y, X(\cdot), \mathbf{z}^*) = \Pr(r = 0|y, X(\cdot), \mathbf{z}^*; \boldsymbol{\theta}_r)$, where $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_r$ contain unknown parameters in the imaging regression $y|(X(\cdot), \mathbf{z})$ and the missing data model $r|(y, X(\cdot), \mathbf{z}^*)$, respectively. Let $\mathcal{D}(\boldsymbol{\theta}_y) \bigotimes \mathcal{D}(\boldsymbol{\theta}_r)$ be the domains of $(\boldsymbol{\theta}_y, \boldsymbol{\theta}_r)$, where $\bigotimes$ is the tensor product of two spaces and $\mathcal{D}(\boldsymbol{\theta}_y)$ and $\mathcal{D}(\boldsymbol{\theta}_r)$ are the domain of $\boldsymbol{\theta}_y$ and that of $\boldsymbol{\theta}_r$, respectively. We will show that the instrumental variable is useful for model identification and estimation in the presence of functional covariates. Moreover, we provide further evidence by including instrumental variables in the simulation study and real data analysis.

Theorem 1. Suppose that there is a nonresponse instrumental covariate u in $\mathbf{z} = (\mathbf{z}^{*T}, u)^T$ such that $\Pr(y|X(\cdot), \mathbf{z})$ depends on u, whereas $\Pr(r = 0|y, X(\cdot), \mathbf{z}) = \Pr(r = 0|y, X(\cdot), \mathbf{z}^*)$ does not depend on u. The model is identifiable under conditions (C1)–(C3) stated below:
(C1) there exists a set $S \subseteq$ the support of $(y, X(\cdot), \mathbf{z}^*)$, such that $\Pr(r = 0|y, X(\cdot), \mathbf{z}^*; \boldsymbol{\theta}_r) \neq 0$ for all $(y, X(\cdot), \mathbf{z}^*) \in S$ and $\boldsymbol{\theta}_r \in \mathcal{D}(\boldsymbol{\theta}_r)$;
(C2) $\Pr(r = 0|y, X(\cdot), \mathbf{z}^*; \boldsymbol{\theta}_{r1}) = \Pr(r = 0|y, X(\cdot), \mathbf{z}^*; \boldsymbol{\theta}_{r2})$ for all $(y, X(\cdot), \mathbf{z}^*) \in S \Longleftrightarrow \boldsymbol{\theta}_{r1} = \boldsymbol{\theta}_{r2}$;
(C3) $p(y|X(\cdot), \mathbf{z}^*, u_1; \boldsymbol{\theta}_{y1})p(y|X(\cdot), \mathbf{z}^*, u_2; \boldsymbol{\theta}_{y2}) = p(y|X(\cdot), \mathbf{z}^*, u_1; \boldsymbol{\theta}_{y2})p(y|X(\cdot), \mathbf{z}^*, u_2; \boldsymbol{\theta}_{y1})$ holds for all (u_1, u_2) and $(y, X(\cdot), \mathbf{z}^*) \in S \Longleftrightarrow \boldsymbol{\theta}_{y1} = \boldsymbol{\theta}_{y2}$.

Theorem 1 extends the theoretical results of Wang, Shao, and Kim (2014) from generalized linear models to the BSOI

model. Specifically, these identifiability conditions (C1)–(C3) can be divided into two parts, including the identifiability of $\Pr(r = 0|y, X(\cdot), \mathbf{z}^*; \boldsymbol{\theta}_r)$ and the identifiability of the likelihood ratio $p(y|X(\cdot), (\mathbf{z}^{*T}, u_1)^T; \boldsymbol{\theta}_y)/p(y|X(\cdot), (\mathbf{z}^{*T}, u_2)^T; \boldsymbol{\theta}_y)$ without missingness. These identification conditions do not need to specify the explicit form of a regression model and are applicable to a large class of model settings, such as semiparametric BSOI regression and the semiparametric imaging exponential tilting model.

As a special case, conditions (C1)–(C3) for the BSOI regression and exponential tilting model can be further clarified to facilitate verification. We need to introduce some notations. Denote the topological support of the random process $X(\cdot)$ by $\mathcal{X}$, which is assumed to be a subset of the quadratically integrable function space $L_2(\mathcal{V})$. Denote the support of $\mathbf{z}^*$ and u by $\mathcal{Z}^*$ and $\mathcal{U}$, which are subsets of $\mathbb{R}^{Q-1}$ and $\mathbb{R}$, respectively. It is assumed that the support of $(X(\cdot), \mathbf{z}^*, u)$ is $\mathcal{X} \bigotimes \mathcal{Z}^* \bigotimes \mathcal{U}$. Given a Hilbert space $H(\langle\cdot,\cdot\rangle)$ and subsets $S_1, S_2 \subseteq H$, we define

$$\mathcal{L}(S_1) = \left\{ \sum_{j=1}^{J} a_j(s_j - \widetilde{s}_j) | s_j, \widetilde{s}_j \in S_1, a_j \in \mathbb{R} \text{ for } j = 1, \ldots, J, J \in \mathbb{Z}_+ \right\}$$

and $\overline{\mathcal{L}}(S_1)$ as the closure of $\mathcal{L}(S_1)$. Moreover, we define

$$\mathcal{L}(S_1, S_2; H)^{\perp} = \{h \in S_2 | \langle h, s\rangle = 0 \text{ for all } s \in \overline{\mathcal{L}}(S_1)\}.$$

Note that $\overline{\mathcal{L}}(S_1)$ is the closed linear span of S_1 if $\mathbf{0} \in S_1$, and $\mathcal{L}(S_1, S_2; H)^{\perp}$ is the perpendicular complement of $\overline{\mathcal{L}}(S_1)$ if $S_2 = H$.

The following proposition holds, and the proof can be found in Appendix B.

Proposition 1. Consider the model specified by (5), (7), and (8) with an instrumental variable as follows:

$$y_i = \alpha + \int_{\mathcal{V}} X_i(v)\beta(v)dv + \boldsymbol{\gamma}_*^T \mathbf{z}_i^* + \gamma_u u_i + \delta_i, \delta_i \sim N(0, \sigma^2),$$
$$\Pr(r_i = 1|y_i, X_i(v), \mathbf{z}_i) = \frac{\exp(\alpha_r + \int_{\mathcal{V}} X_i(v)\beta_r(v)dv + \boldsymbol{\gamma}_r^T \mathbf{z}_i^* + \phi y_i)}{1 + \exp(\alpha_r + \int_{\mathcal{V}} X_i(v)\beta_r(v)dv + \boldsymbol{\gamma}_r^T \mathbf{z}_i^* + \phi y_i)}.$$

The following conditions are sufficient conditions of (C1)–(C3):
(i) there exists an $\epsilon > 0$ such that $|\gamma_u| \geq \epsilon$;
(ii) $\min(\|\mathcal{X}\|_0, \|\mathcal{Z}^*\|_0, \|\mathcal{U}\|_0) \geq 2$, where $\|\cdot\|_0$ is the number of elements.
(iii) $\mathcal{L}(\mathcal{X}, \mathcal{D}(\beta(\cdot)); L_2(\mathcal{V}))^{\perp}$, $\mathcal{L}(\mathcal{X}, \mathcal{D}(\beta_r(\cdot)); L_2(\mathcal{V}))^{\perp}$, $\mathcal{L}(\mathcal{Z}^*, \mathcal{D}(\boldsymbol{\gamma}_*); \mathbb{R}^{Q-1})^{\perp}$, and $\mathcal{L}(\mathcal{Z}^*, \mathcal{D}(\boldsymbol{\gamma}_r); \mathbb{R}^{Q-1})^{\perp}$ are all zero, where $\mathcal{D}(\beta(\cdot))$, $\mathcal{D}(\beta_r(\cdot))$, $\mathcal{D}(\boldsymbol{\gamma}_*)$ and $\mathcal{D}(\boldsymbol{\gamma}_r)$ are the domains of $\beta(\cdot)$, $\beta_r(\cdot)$, $\boldsymbol{\gamma}_*$, and $\boldsymbol{\gamma}_r$, respectively.

Conditions (i), (ii), and (iii) in Proposition 1 can be easily satisfied, as demonstrated in the following example.

Example 2. Let $\mathcal{V} = [0, T]$ for some $T > 0$. $X(\cdot)$ is a continuous stochastic process on $\mathcal{V}$ with topological support $\mathcal{X} = \{X(\cdot) \in C[0, T] | X(0) = 0\}$, for example, the one-dimensional Wiener process with nonsingular covariance matrix. The $\beta(\cdot)$ and $\beta_r(\cdot) \in C[0, T]$ are continuous functions on $[0, T]$, indicating $\mathcal{D}(\beta_r(\cdot)) = \mathcal{D}(\beta(\cdot)) = C[0, T]$. It follows from the fact that $\mathcal{X}$ is dense in $\widetilde{\mathcal{X}} = \{X(\cdot) \in L_2[0, T] | X(0) = 0\}$ that we have

$$\mathcal{L}(\mathcal{X}, \mathcal{D}(\beta(\cdot)); L_2(\mathcal{V}))^{\perp} = \mathcal{L}(\mathcal{X}, \mathcal{D}(\beta_r(\cdot)); L_2(\mathcal{V}))^{\perp} = \mathcal{L}(\widetilde{\mathcal{X}}, C[0, T]; L_2([0, T]))^{\perp} = 0.$$

Suppose $\mathcal{Z}^* = \mathbb{R}^{Q-1}$, $\mathcal{D}(\boldsymbol{\gamma}_*) = \mathbb{R}^{Q-1}$, and $\mathcal{D}(\boldsymbol{\gamma}_r) = \mathbb{R}^{Q-1}$. Then, we have

$$\mathcal{L}(\mathcal{Z}^*, \mathcal{D}(\boldsymbol{\gamma}_*); \mathbb{R}^{Q-1})^{\perp} = \mathcal{L}(\mathcal{Z}^*, \mathcal{D}(\boldsymbol{\gamma}_r); \mathbb{R}^{Q-1})^{\perp} = \mathcal{L}(\mathbb{R}^{Q-1}, \mathbb{R}^{Q-1}; \mathbb{R}^{Q-1})^{\perp} = 0.$$

The key condition $|\gamma_u| \geq \epsilon$ guarantees that the dependency of the response y on the instrumental variable u exists.

3.2. Bayesian Inference

Let $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{mis}}$ be the vectors of observed and missing responses, respectively; $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$; $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ be the matrix of observed covariates; and $\mathbf{r} = (r_1, \ldots, r_n)^T$ be a vector of missingness indicators. We also denote $\boldsymbol{\theta}_y = (\alpha, \beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_Q, \sigma^2)^T$ as a vector that includes all the unknown parameters involved in the BSOI regression model (6), $\boldsymbol{\theta}_r = (\alpha_r, \beta_{r1}, \ldots, \beta_{rK}, \gamma_{r1}, \ldots, \gamma_{r,Q-1})^T$ as a vector that includes all the unknown parameters involved in missing data model (9), and $\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_r)$. Recall that $\mathbf{X}$ is the matrix of imaging observations.

After performing FPCA on the imaging observations $X_i(v)$, we can consider the eigenscores ξ_{ik} as known covariates in the regression model. Consequently, the BSOI regression model is reduced to a conventional linear model and is readily analyzed under the Bayesian framework. We then assign conjugate priors for the parameters in $\boldsymbol{\theta}_y$ as follows:

$$\begin{aligned} &p(\alpha) \overset{D}{=} \mathrm{N}(\alpha_0, \sigma_{\alpha 0}^2); \quad p(\beta_k) \overset{D}{=} \mathrm{N}(\beta_{k0}, \sigma_{\beta_k 0}^2), \quad k = 1, \ldots, K; \\ &p(\gamma_q) \overset{D}{=} \mathrm{N}(\gamma_{q0}, \sigma_{\gamma_q 0}^2), \quad q = 1, \ldots, Q; \quad \text{and} \\ &p(\sigma^{-2}) \overset{D}{=} \mathrm{Gamma}(a_{\sigma 0}, b_{\sigma 0}), \end{aligned} \tag{12}$$

where α_0, $\sigma_{\alpha 0}^2$, β_{k0}, $\sigma_{\beta_k 0}^2$, γ_{q0}, $\sigma_{\gamma_q 0}^2$, $a_{\sigma 0}$, and $b_{\sigma 0}$ are the hyperparameters, and their values are prespecified according to the information from historical analyses or prior knowledge. For the parameter vector $\boldsymbol{\theta}_r$ involved in (9), we assign the prior distribution as follows:

$$p(\boldsymbol{\theta}_r) \overset{D}{=} \mathrm{N}(\boldsymbol{\theta}_{r0}, \boldsymbol{\Sigma}_{r0}), \tag{13}$$

where $\boldsymbol{\theta}_{r0}$ and $\boldsymbol{\Sigma}_{r0}$ are prespecified hyperparameters.

The Bayesian estimate of $\boldsymbol{\theta}$ can be obtained through sampling from $p(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}, \mathbf{Z}, \mathbf{X})$. Owing to the existence of nonresponses, this posterior distribution involves a high-dimensional integral and is therefore intractable. With the use of a data augmentation technique (Tanner and Wong 1987), we work on the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \mathbf{r}, \mathbf{Z}, \mathbf{X})$. MCMC methods, such as the Gibbs sampler and the Metropolis–Hastings algorithm, are used iteratively to sample (I) $\mathbf{y}_{\text{mis}}$ from $p(\mathbf{y}_{\text{mis}}|\boldsymbol{\theta}, \mathbf{y}_{\text{obs}}, \mathbf{r}, \mathbf{Z}, \mathbf{X})$, (II) $\boldsymbol{\theta}_y$ from $p(\boldsymbol{\theta}_y|\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}, \mathbf{Z}, \mathbf{X})$, and (III) $\boldsymbol{\theta}_r$ from $p(\boldsymbol{\theta}_r|\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}}, \mathbf{r}, \mathbf{Z}, \mathbf{X})$. The conditional

distributions and technical details are provided in Appendix C. The computer program is written in the R language with the aid of RcppArmadillo package (Eddelbuettel and Sanderson 2014) for speeding up loops in the code, and the main functions are summarized in the R package "BSOINN."

4. Simulation Study

In this section, two simulations are conducted to evaluate the empirical performance of the proposed method. In Simulation 1, we assess the Bayesian estimation of the BSOI regression with nonignorable, ignorable, and fully observable data. In Simulation 2, we further consider several numerical studies that evaluate effectiveness of the instrumental variable in improving the model identification and estimation of our proposed method.

4.1. Simulation 1

We consider 2D $V_1 \times V_2$ imaging data with $V_1 = V_2 = 300$, resulting in a $\mathbf{X}_i$ with a length of $V = 90{,}000$. The data are simulated from the following model:

$$y_i = \int_{\mathcal{V}} X_i(v)\beta(v)dv + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 u_i + \delta_i \qquad (14)$$

with $X_i(v) = \sum_{k=1}^{3} \xi_{ik}\psi_k(v)$ and $\beta(v) = \sum_{k=1}^{3} \beta_k\psi_k(v)$, where $\xi_{ik} \sim N(0, 0.5^{k-1})$, $\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = -1$, $\gamma_1 = 1.5$, $\gamma_2 = -1$, $\gamma_3 = 0.5$, and $\delta_i \sim N(0, 1)$. The covariates z_{i1} and z_{i2} are independently generated from $U(0, 1)$ and $N(0, 1)$, respectively. Moreover, $U(0, 1)$ denotes the uniform distribution in $[0, 1]$, and u_i denotes a binary instrumental variable with equal probabilities taking the values of 1 and -1. The eigenimages ψ_k are presented in the first row of Figure 2. They can be regarded as two-dimensional grayscale images with pixel intensities on a [0,1] scale, where the black pixels are set as 1 and white pixels are set as 0. Such a method of generating imaging data from eigenimages can also be found in studies conducted by Zipunnikov et al. (2011a, 2011b).

The two missing mechanisms are considered as follows.

Mechanism 1 (Nonignorable): The missing data are generated on the basis of an exponential tilting model with imaging predictors as follows:

$$\text{logit}(\pi_i) = \int_{\mathcal{V}} X_i(v)\beta_r(v)dv + \gamma_{r1} z_{i1} + \gamma_{r2} z_{i2} + \phi y_i, \quad (15)$$

where $\beta_r(v) = \sum_{k=1}^{3} \beta_{rk}\psi_k(v)$, $\beta_{r1} = -1$, $\beta_{r2} = 0.5$, $\beta_{r3} = 0.5$, $\gamma_{r1} = -0.7$, $\gamma_{r2} = -0.7$, and $\phi = -1.2$. The overall missing proportion is approximately 40%.

Mechanism 2 (Ignorable): The missing data are generated on the basis of a logistic regression model with imaging predictors as follows:

$$\text{logit}(\pi_i) = \int_{\mathcal{V}} X_i(v)\beta_r(v)dv + \gamma_{r1} z_{i1} + \gamma_{r2} z_{i2}, \qquad (16)$$

where $\beta_{r1} = -1$, $\beta_{r2} = 0.5$, $\beta_{r3} = 0.5$, $\gamma_{r1} = -1$, and $\gamma_{r2} = 0.7$. Mechanism 2 is in fact a special case of Mechanism 1, the tilting parameter ϕ of which is set to zero. The overall missing proportion is also approximately 40%.

In the above designs, Mechanism 1 is used to evaluate whether the proposed method can accurately retrieve the information of nonignorable missingness, whereas Mechanism 2 is employed to investigate the implication of overspecification of a missing data model. Thus, we consider the following models in the data analysis:

BSOI-NN: BSOI regression (14) and a nonignorable missing model (8) specified as model (15).
BSOI-IN: BSOI regression (14) and an ignorable missing model (8) specified as model (16).
BSOI-Full: BSOI regression (14) with fully observed data.

Notably, the BSOI-Full model assumes that the values of the missing observations are known. Thus, it can be considered as an oracle model, and its corresponding estimation results can be considered as a benchmark for comparison.

Regarding the prior distributions in (12) and (13), we set vague prior inputs as follows: $\alpha_0 = 0, \sigma^2_{\alpha 0} = 10$, $\beta_{k0} = 0, \sigma^2_{\beta_k 0} = 10$, $\gamma_{q0} = 0, \sigma^2_{\gamma_q 0} = 10$, and $a_{\sigma 0} = 9, b_{\sigma 0} = 3$, $\boldsymbol{\theta}_{r0} = \mathbf{0}$, and $\boldsymbol{\Sigma}_{r0} = \mathbf{I}$, where $\mathbf{0}$ and $\mathbf{I}$ denote the zero vector and identity matrices with appropriate dimensions, respectively. We assess the convergence of the MCMC algorithm using three parallel sequences with well-separated starting values. The MCMC algorithm converges within 4000 iterations. Thus, we collect 6000 observations after 4000 burn-in iterations to conduct Bayesian inference.

In this simulation, three sample sizes, $n = 100$, $n = 500$, and $n = 1000$, are considered. A total of 100 replicated datasets are generated under each sample size and missing mechanism. We first look at the performance of FPCA on the imaging data. The estimated eigenimages and eigenscores with the moderate sample size $n = 500$ are depicted in Figures 2 and 3. In Figure 2, the first and second rows from left to right provide the true eigenimages and the means of their estimates, respectively, whereas the last row displays the means and the 5th and 95th percentiles of the estimated eigenimages. The estimated eigenimages are normalized through $(\hat{\psi}_k(v) - \min_v \hat{\psi}_k(v))/(\max_v \hat{\psi}_k(v) - \min_v \hat{\psi}_k(v))$ to obtain grayscale images with pixel values in the [0,1] interval. As shown in Figure 2, the means of the estimated eigenimages perfectly recover the spatial configuration. The small distortions from the true eigenimages to their 5th and 95th percentiles reflect the good performance of the estimated eigenimages. Figure 3 reports the estimation result of the eigenscores. Both the Q-Q plot and the boxplots show that the estimated eigenscores are very close to their true values.

To evaluate the finite sample performance of the parameter estimates, we compute the bias (BIAS) and root mean squared error (RMS) between the Bayesian estimates of the unknown parameters and their true population values. Table 1 (upper panel) presents the estimation results obtained from the 100 replicated datasets generated under Mechanism 1. As expected, the estimates under BSOI-Full (oracle case) have excellent performance with small BIAS and RMS values. The estimates under BSOI-NN (true mechanism) are not as good as those of BSOI-Full but are still satisfactory. The estimates under both cases improve when the sample size increases. Meanwhile, the estimates under BSOI-IN (oversim-

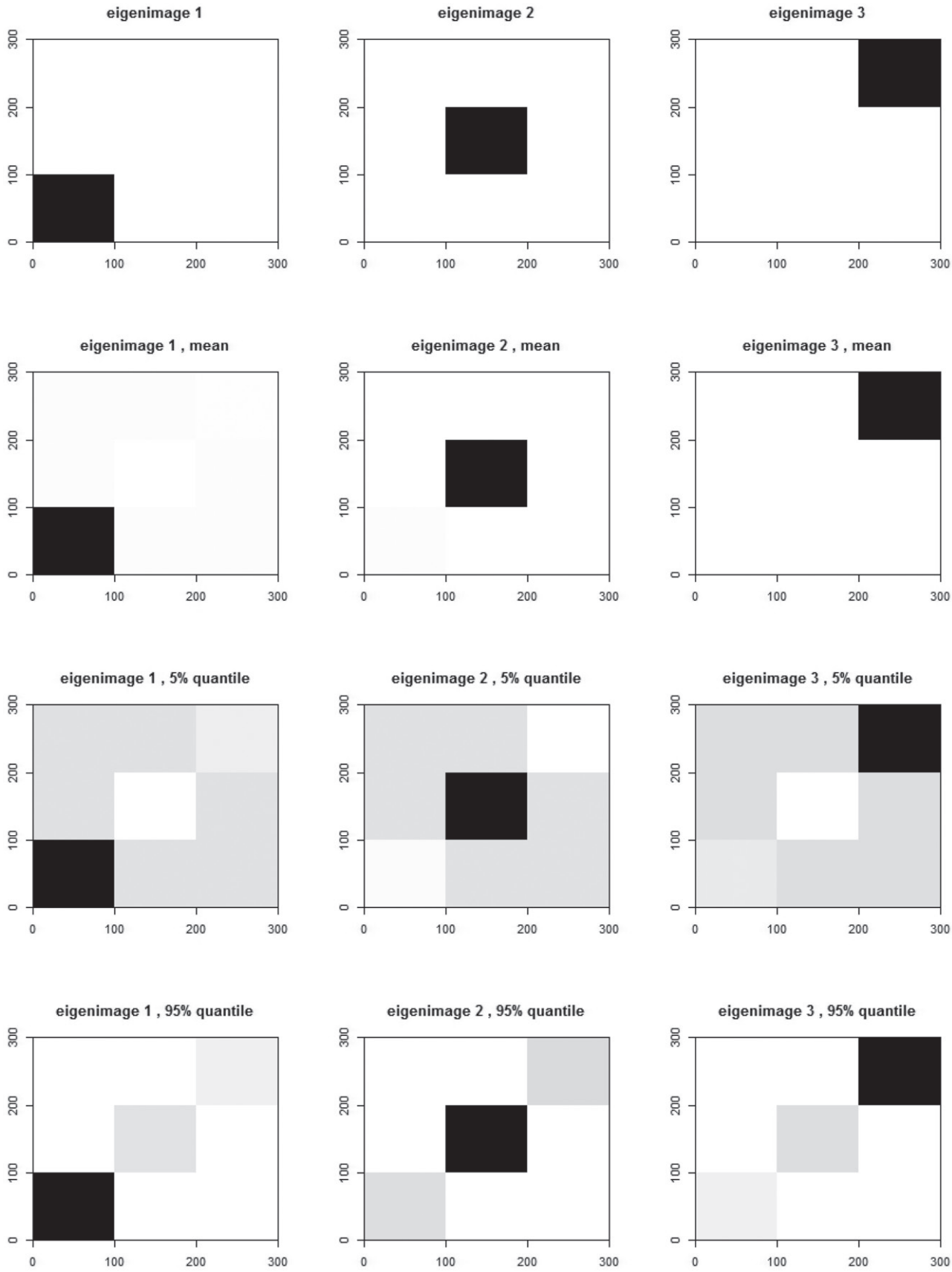

Figure 2. True (top row), estimated mean (2nd row), 5th pointwise percentile (3rd row), and 95th pointwise percentile (bottom row) grayscale eigenimages in Simulation 1 with the sample size $n = 500$.

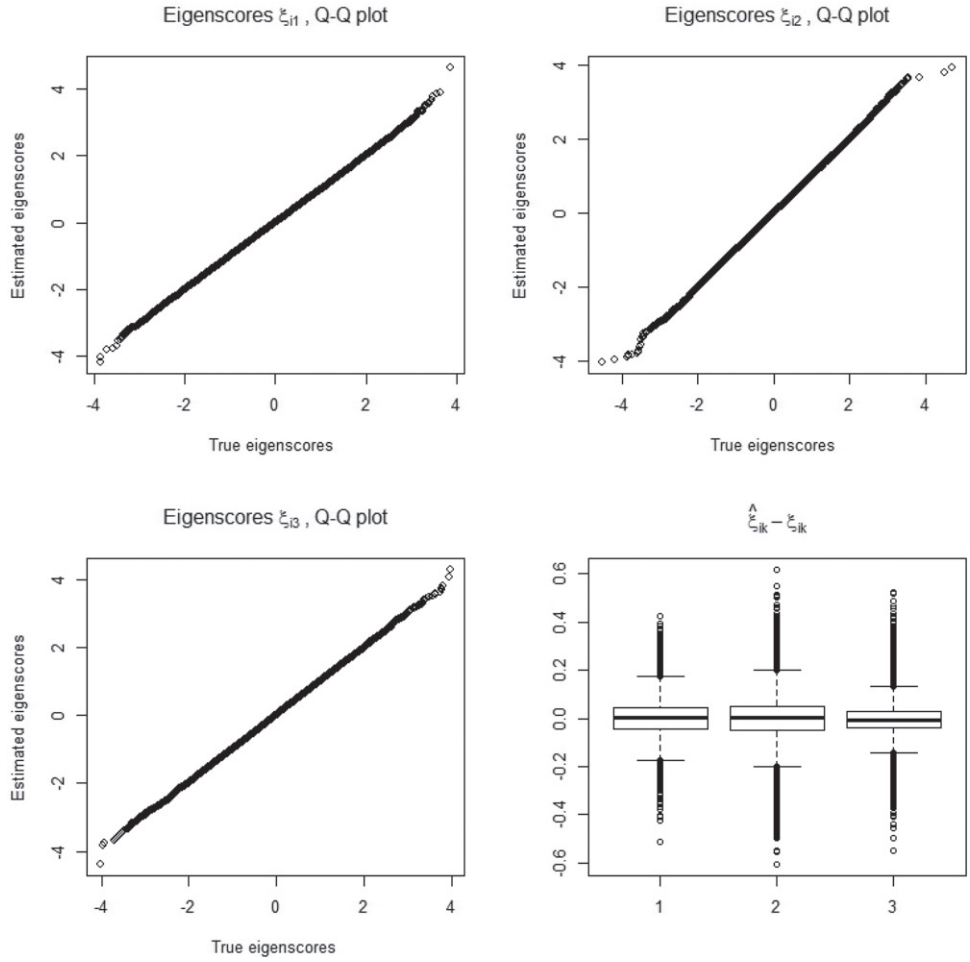

Figure 3. Q-Q plots between the true and estimated eigenscores and the boxplots of $\hat{\xi}_{ik} - \xi_{ik}$ in Simulation 1 with the sample size $n = 500$.

plified mechanism) perform unsatisfactorily with large BIAS and RMS values, indicating that ignoring an nonignorable missing mechanism can lead to seriously biased estimation results.

For comparison, the estimation results obtained on the basis of the 100 replicated datasets generated under Mechanism 2 are presented in Table 1 (lower panel). The performance of the estimates under BSOI-Full (oracle case) is excellent, whereas the performances of the estimates under BSOI-IN (true mechanism) and BSOI-NN (overspecified mechanism) are satisfactory and comparable. Thus, the BSOI-NN procedure does not distort the parameter estimates of BSOI regression even when the true missing mechanism is ignorable.

The sensitivity of Bayesian estimates to prior inputs is investigated by introducing certain disturbances to the hyperparameters. For example, we reanalyze the previous analysis by setting $\sigma^2_{\alpha 0} = 100$, $\sigma^2_{\beta_k 0} = 100$, $\sigma^2_{\gamma_q 0} = 100$, and some ad hoc disturbances to other hyperparameters. The obtained results are similar and are not reported here.

4.2. Simulation 2

4.2.1. Part 1

In this part, we evaluate the effectiveness of the instrumental variable in improving model identification and estimation. Specifically, we generate datasets using the same setting of model (14) in Simulation 1 except that the instrumental variable u_i is removed. We generate 100 replicated datasets of different sample sizes under Mechanisms 1 and 2 that are specified by (15) and (16). The proposed BSOI-NN model, which consists of a BSOI regression (14) without u_i and a nonignorable missing model (8) specified as the model of (15), is then applied to analyze the newly generated datasets. The estimation results are compared with those obtained in Simulation 1, which implements the BOSI-NN method on datasets with an instrumental variable (see Table 2). Notably, for the sake of fairness, the mean value of the instrumental variable u_i is intentionally set to 0 in Simulation 1 to ensure that the instrumental variable does not affect the overall missing rate of the responses.

As shown in Table 2, the estimation with an instrumental variable outperforms that without the instrumental variable. In particular, for the unknown parameters involved in the exponential tilting model, their estimates with an instrumental variable uniformly improve as the sample size increases. However, such improvement is not achieved when the instrumental variable is excluded, and the estimates of several parameters, such as β_{r2}, γ_{r1}, γ_{r2}, and ϕ, have undesirable results as the sample size increases from 500 to 1000. Thus, the instrumental variable facilitates model identification and estimation.

Table 1. Bayesian parameter estimates in Simulation 1.

	$n = 100$						$n = 500$						$n = 1000$					
	BSOI-Full		BSOI-IN		BSOI-NN		BSOI-Full		BSOI-IN		BSOI-NN		BSOI-Full		BSOI-IN		BSOI-NN	
Para	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS
Mechanism 1																		
$\alpha = 0$	−0.011	0.183	0.526	0.598	0.041	0.306	−0.007	0.083	−0.509	0.522	−0.012	0.157	−0.009	0.060	0.502	0.508	−0.018	0.117
$\beta_1 = 0.5$	−0.011	0.213	−0.182	0.297	−0.002	0.226	0.008	0.076	−0.170	0.191	0.009	0.095	0.003	0.055	−0.170	0.180	0.010	0.071
$\beta_2 = 1$	0.001	0.220	−0.072	0.255	−0.021	0.245	−0.019	0.094	−0.095	0.130	−0.021	0.104	−0.002	0.066	−0.072	0.099	0.004	0.077
$\beta_3 = -1$	0.046	0.281	0.254	0.393	0.025	0.322	0.005	0.109	−0.202	0.237	0.013	0.136	−0.001	0.071	0.173	0.191	−0.016	0.092
$\gamma_1 = 1.5$	0.041	0.324	−0.299	0.557	−0.001	0.468	0.027	0.149	−0.269	0.317	−0.022	0.193	0.007	0.098	−0.283	0.307	0.005	0.135
$\gamma_2 = -1$	−0.002	0.099	0.053	0.131	0.001	0.125	0.000	0.045	0.064	0.080	0.009	0.052	−0.002	0.035	0.051	0.064	−0.005	0.039
$\gamma_3 = 0.5$	−0.024	0.095	−0.095	0.153	−0.026	0.125	0.003	0.038	−0.063	0.084	0.008	0.052	0.001	0.030	−0.068	0.079	0.003	0.038
$\sigma_\delta^2 = 1$	−0.093	0.148	−0.252	0.284	−0.114	0.199	−0.024	0.067	−0.162	0.175	−0.020	0.099	−0.016	0.041	−0.150	0.156	−0.008	0.077
$\alpha_r = 0.5$					−0.128	0.468					−0.057	0.279					−0.030	0.195
$\beta_{r1} = -1$					−0.092	0.413					−0.029	0.163					−0.014	0.125
$\beta_{r2} = 0.5$					−0.058	0.550					0.071	0.314					0.062	0.243
$\beta_{r3} = 0.5$					0.037	0.552					−0.016	0.353					−0.070	0.253
$\gamma_{r1} = -0.7$					0.118	0.603					0.094	0.501					0.039	0.350
$\gamma_{r2} = -0.7$					0.020	0.423					−0.070	0.293					−0.061	0.235
$\phi = -1.2$					−0.141	0.405					−0.134	0.323					−0.086	0.243
Mechanism 2																		
$\alpha = 0$	−0.002	0.192	0.013	0.287	−0.074	0.349	−0.006	0.087	−0.029	0.120	−0.007	0.168	−0.010	0.059	−0.001	0.087	−0.021	0.104
$\beta_1 = 0.5$	−0.008	0.164	−0.005	0.212	0.024	0.209	−0.006	0.086	−0.001	0.101	0.015	0.111	0.004	0.055	0.002	0.065	0.009	0.065
$\beta_2 = 1$	−0.006	0.205	−0.005	0.276	−0.020	0.226	0.000	0.088	−0.011	0.101	−0.017	0.102	−0.002	0.066	−0.001	0.081	−0.004	0.079
$\beta_3 = -1$	0.018	0.270	0.039	0.306	0.025	0.308	−0.002	0.109	−0.007	0.133	−0.014	0.134	−0.001	0.070	−0.001	0.093	−0.004	0.093
$\gamma_1 = 1.5$	−0.013	0.327	−0.040	0.462	−0.019	0.470	0.004	0.144	0.023	0.195	0.039	0.195	0.008	0.097	−0.008	0.127	−0.001	0.129
$\gamma_2 = -1$	0.034	0.101	0.026	0.154	0.003	0.163	0.002	0.047	0.003	0.060	−0.007	0.062	−0.002	0.035	−0.006	0.044	−0.010	0.045
$\gamma_3 = 0.5$	0.001	0.105	−0.002	0.157	−0.000	0.157	0.001	0.045	−0.007	0.057	−0.006	0.056	0.001	0.030	−0.001	0.037	−0.001	0.037
$\sigma_\delta^2 = 1$	−0.114	0.170	−0.187	0.241	−0.149	0.222	−0.023	0.063	−0.027	0.084	−0.012	0.083	−0.016	0.041	−0.025	0.055	−0.018	0.053
$\alpha_r = 0$					−0.295	0.448					−0.035	0.222					−0.042	0.155
$\beta_{r1} = -1$					−0.008	0.364					−0.001	0.162					0.005	0.114
$\beta_{r2} = 0.5$					0.241	0.539					0.103	0.261					0.052	0.187
$\beta_{r3} = 0.5$					−0.143	0.556					−0.091	0.271					−0.015	0.198
$\gamma_{r1} = -1$					0.585	0.841					0.111	0.496					0.086	0.330
$\gamma_{r2} = 0.7$					−0.101	0.407					−0.063	0.256					−0.030	0.159
$\phi = 0$					−0.209	0.408					−0.090	0.232					−0.042	0.136

4.2.2. Part 2

In Simulation 1, we assume that the coefficient image $\beta(\cdot)$ can be well represented by the eigenimages of $X(\cdot)$. This part aims to evaluate the performance of the proposed method when $\beta(\cdot)$ is not directly generated from the eigenimages. By fixing all the other settings exactly the same as Mechanism 1 of Simulation 1, we consider two cases of true $\beta(\cdot)$ as shown in Figure 4, which cannot be directly generated by the eigenimages (see Figure 2) of $X(\cdot)$. We consider a moderate sample size $n = 500$ in this simulation and generate 100 replicated datasets for each case. The three methods, BSOI-NN, BSOI-IN, and BSOI-Full, are utilized for data analysis. The estimation results are shown in Figure 4 and Table 3, with the estimation accuracy of $\beta(\cdot)$ being assessed through the following RMS measurement:

$$\text{RMS} = \sqrt{\frac{1}{100}\sum_{l=1}^{100}\left[\frac{1}{V}\sum_{v=1}^{V}\{\hat{\beta}_l(v) - \beta(v)\}^2\right]},$$

where $\hat{\beta}_l(v)$ is the estimate of $\beta(v)$ at the vth voxel in the lth replication.

As depicted in Figure 4 and Table 3, we find that the BSOI-Full method with full observations still performs the best in both cases on reproducing $\beta(\cdot)$ and other parameters, and the proposed BSOI-NN method exhibits a comparable performance. In contrast, BSOI-IN method performs unsatisfactorily in both cases. Nonetheless, if $\beta(\cdot)$ cannot be well represented by the eigenimages, none of the three methods is able to recover $\beta(\cdot)$ accurately.

The considered scalar-on-image regression inherently requires certain structural assumptions on the coefficient image $\beta(\cdot)$ because the dimension of imaging covariates is much larger than the sample size (Happ, Greven, and Schmid 2018; Kang, Reich, and Staicu 2018; Wang and Zhu 2017, among others). A common assumption is that $\beta(\cdot)$ is a linear combination of the leading eigenimages of $X(\cdot)$, which we considered in the proposed BSOI model. This assumption seems plausible because areas with high variation in the imaging observations are likely to be relevant to the response values (Happ, Greven, and Schmid 2018). Recent advancements on scalar-on-image regression assume spatial sparsity and smoothness on voxels of $\beta(\cdot)$ to identify $\beta(\cdot)$ (Wang and Zhu 2017; Kang, Reich, and Staicu 2018). However, such voxelwise methods are extremely time consuming, especially for managing the current ultrahigh dimensional brain images. Nevertheless, none of the aforementioned assumptions fits every real situation. The existing literature (Happ, Greven, and Schmid 2018; Kang, Reich, and Staicu 2018) shows that the estimated $\beta(\cdot)$s based on different assumptions may vary but should have some common features in the detected regions. Specifically, regions of $\beta(\cdot)$ estimated to be positive (negative) effects under one assumption tend to be positive (negative) under other assumptions. This

Table 2. Bayesian parameter estimates in Part 1 of Simulation 2.

	$n = 100$				$n = 500$				$n = 1000$			
	Without		With		Without		With		Without		With	
Para	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS
Mechanism 1												
$\alpha = 0$	0.124	0.356	0.041	0.306	−0.036	0.180	−0.012	0.157	0.026	0.173	−0.018	0.117
$\beta_1 = 0.5$	−0.092	0.248	−0.002	0.226	−0.012	0.116	0.009	0.095	−0.018	0.092	0.010	0.071
$\beta_2 = 1$	−0.034	0.259	−0.021	0.245	0.001	0.109	−0.021	0.104	−0.002	0.075	0.004	0.077
$\beta_3 = -1$	0.039	0.347	0.025	0.322	0.044	0.163	0.013	0.136	0.008	0.104	−0.016	0.092
$\gamma_1 = 1.5$	−0.094	0.531	−0.001	0.468	−0.013	0.196	−0.022	0.193	−0.003	0.163	0.005	0.135
$\gamma_2 = -1$	0.033	0.141	0.001	0.125	0.004	0.063	0.009	0.052	0.001	0.043	−0.005	0.039
$\gamma_3 = 0.5$			−0.026	0.125			0.008	0.052			0.003	0.038
$\sigma_\delta^2 = 1$	−0.149	0.231	−0.114	0.199	−0.019	0.108	−0.020	0.099	−0.012	0.104	−0.008	0.077
$\alpha_r = 0.5$	−0.296	0.527	−0.128	0.468	−0.097	0.288	−0.057	0.279	−0.031	0.207	−0.030	0.195
$\beta_{r1} = -1$	−0.165	0.487	−0.092	0.413	−0.061	0.207	−0.029	0.163	−0.040	0.137	−0.014	0.125
$\beta_{r2} = 0.5$	−0.261	0.593	−0.058	0.550	−0.056	0.354	0.071	0.314	−0.019	0.363	0.062	0.243
$\beta_{r3} = 0.5$	0.146	0.650	0.037	0.552	0.096	0.375	−0.016	0.353	0.027	0.319	−0.070	0.253
$\gamma_{r1} = -0.7$	0.121	0.612	0.118	0.603	−0.054	0.521	0.094	0.501	−0.091	0.521	0.039	0.350
$\gamma_{r2} = -0.7$	0.212	0.445	0.020	0.423	0.063	0.357	−0.070	0.293	0.019	0.379	−0.061	0.235
$\phi = -1.2$	0.064	0.410	−0.141	0.405	0.030	0.378	−0.134	0.323	0.020	0.413	−0.086	0.243
Mechanism 2												
$\alpha = 0$	−0.212	0.407	−0.074	0.349	−0.169	0.242	−0.007	0.168	−0.132	0.231	−0.021	0.104
$\beta_1 = 0.5$	0.088	0.229	0.024	0.209	0.047	0.111	0.015	0.111	0.035	0.099	0.009	0.065
$\beta_2 = 1$	−0.036	0.245	−0.020	0.226	−0.012	0.102	−0.017	0.102	−0.018	0.091	−0.004	0.079
$\beta_3 = -1$	−0.027	0.347	0.025	0.308	−0.019	0.143	−0.014	0.134	−0.026	0.110	−0.004	0.093
$\gamma_1 = 1.5$	0.065	0.496	−0.019	0.470	0.093	0.217	0.039	0.195	0.063	0.177	−0.001	0.129
$\gamma_2 = -1$	−0.052	0.180	0.003	0.163	−0.032	0.075	−0.007	0.062	−0.030	0.075	−0.010	0.045
$\gamma_3 = 0.5$			−0.000	0.157			−0.006	0.056			−0.001	0.037
$\sigma_\delta^2 = 1$	−0.092	0.235	−0.149	0.222	0.025	0.104	−0.012	0.083	0.044	0.088	−0.018	0.053
$\alpha_r = 0$	−0.240	0.459	−0.295	0.448	−0.174	0.276	−0.035	0.222	−0.159	0.231	−0.042	0.155
$\beta_{r1} = -1$	0.145	0.353	−0.008	0.364	0.140	0.250	−0.001	0.162	0.108	0.245	0.005	0.114
$\beta_{r2} = 0.5$	0.413	0.576	0.241	0.539	0.361	0.475	0.103	0.261	0.258	0.474	0.052	0.187
$\beta_{r3} = 0.5$	−0.423	0.694	−0.143	0.556	−0.340	0.490	−0.091	0.271	−0.269	0.473	−0.015	0.198
$\gamma_{r1} = -1$	0.655	0.859	0.585	0.841	0.484	0.708	0.111	0.496	0.378	0.653	0.086	0.330
$\gamma_{r2} = 0.7$	−0.407	0.554	−0.101	0.407	−0.317	0.438	−0.063	0.256	−0.253	0.460	−0.030	0.159
$\phi = 0$	−0.462	0.525	−0.209	0.408	−0.351	0.448	−0.090	0.232	−0.264	0.450	−0.042	0.136

suggests a remedy to check the estimation results using a different assumption in the real data analysis. It is worth noting that recovering $\beta(\cdot)$ accurately under different situations for imaging regression is still an active research area, and we can adopt advanced techniques of estimating $\beta(\cdot)$ to improve the proposed BSOI-NN method. We acknowledge this limitation in the final discussion and will investigate it further in our future research.

4.2.3. Part 3

This part evaluates the out-of-sample prediction performance of the proposed BSOI-NN method and compares its performance with those of several other models. In addition to BSOI-Full, BSOI-IN, and BSOI-NN, we also consider two scalar-on-image regression models that were recently developed for fully observed imaging datasets, namely, the scalar-on-image regression model via the soft-thresholded Gaussian process (STGP, Kang, Reich, and Staicu 2018) and the scalar-on-image regression model via total variation (TV, Wang and Zhu 2017). The STGP method models the coefficient image $\beta(\cdot)$ through soft-thresholding of a latent Gaussian process, which not only ensures a gradual transition between the zero and nonzero effects of neighboring locations but also provides large support over the class of spatially varying regression coefficient images. In contrast, the TV method assumes that $\beta(\cdot)$ belongs to the space of bounded total variation, which explicitly accounts for the common piecewise smooth nature of imaging data. The STGP approach has been implemented in the R package "STGP," and the default settings therein are used in this study. The TV approach has been implemented through MATLAB with suggested settings being applied. The target of considering these two alternatives is to assess how much improvement the proposed method can achieve by considering nonignorable missing responses compared with existing scalar-on-image regression methods that disregard missingness.

We reuse the 100 replicated datasets generated in Simulation 1 for Mechanism 1 with a moderate sample size $n = 500$. For each replicated dataset, we evenly distribute it into a training subset and a testing subset, ensuring that both subsets exhibit the same missing rate. The training subsets are utilized to fit the imaging regression model, and the testing subsets are used to evaluate the out-of-sample prediction accuracy. For BSOI-Full method, we again substitute the missing responses with their true values for the training subsets and use the full data to fit the model. The prediction performance of BSOI-Full serves as a benchmark in the comparison. For BSOI-IN and BSOI-NN methods, we use the training subsets with nonresponses to fit models. For STGP and TV methods, we discard the samples with missing responses and use the remaining observations to fit models. The parameter estimates obtained from different methods are used to obtain predicted responses. Following Li et al. (2018), we calculate the Pearson correlation between the predicted responses and their true values in testing subsets as the measurement of prediction accuracy. The obtained

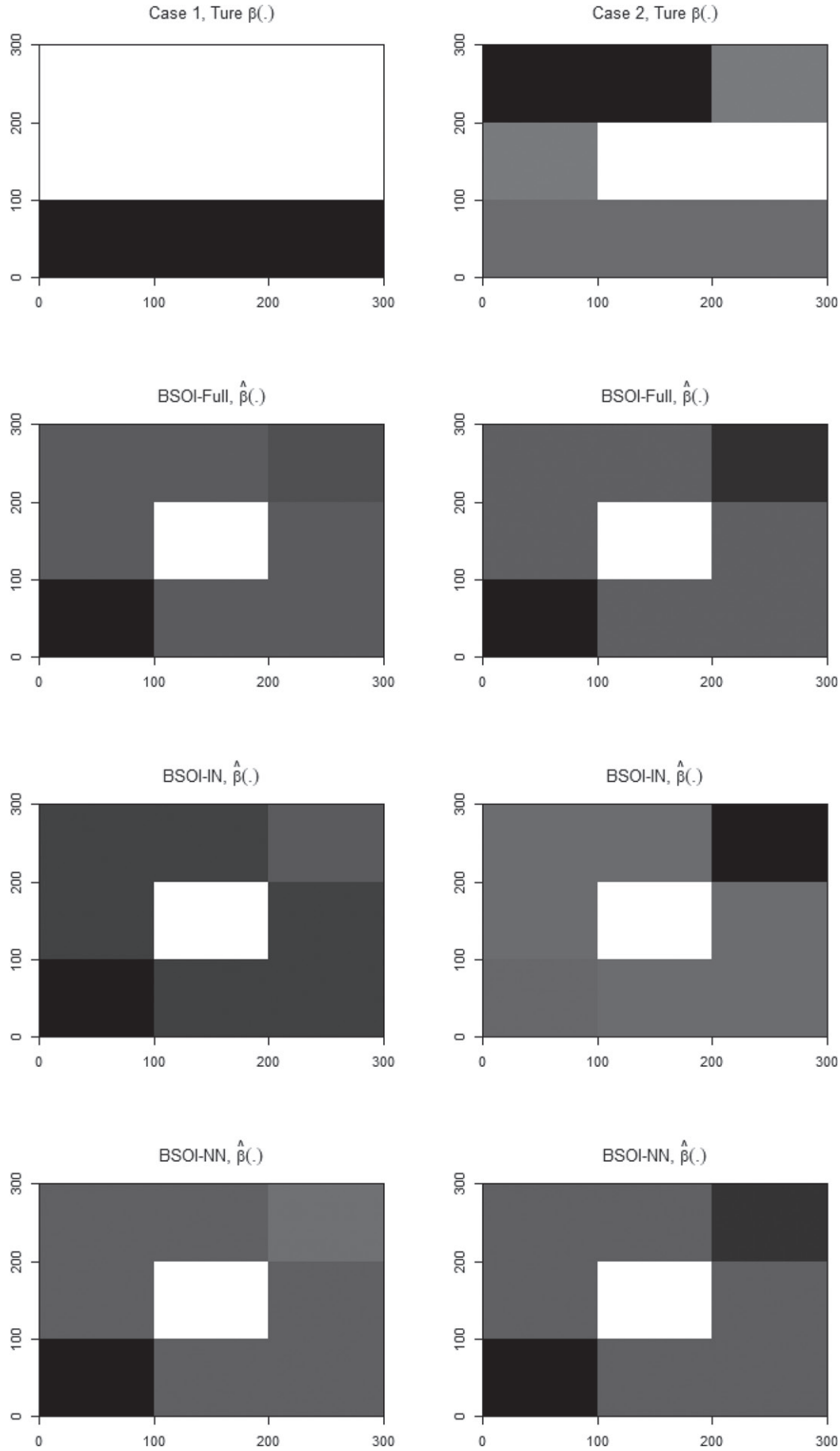

Figure 4. Coefficient image $\beta(\cdot)$ in Part 2 of Simulation 2, where the first (second) column is for case 1 (case 2).

Table 3. Bayesian parameter estimates in Part 2 of Simulation 2 with $n = 500$.

	Case 1						Case 2					
	BSOI-Full		BSOI-IN		BSOI-NN		BSOI-Full		BSOI-IN		BSOI-NN	
Para	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS	BIAS	RMS
$\alpha = 0$	−0.012	0.092	0.486	0.503	−0.040	0.152	−0.006	0.092	0.506	0.521	−0.028	0.168
$\beta(\cdot)$		0.275		0.286		0.276		0.264		0.275		0.266
$\gamma_1 = 1.5$	0.026	0.161	−0.254	0.333	0.048	0.218	0.002	0.165	−0.282	0.344	0.039	0.227
$\gamma_2 = -1$	0.004	0.044	0.057	0.076	−0.003	0.052	0.004	0.043	0.065	0.080	0.002	0.051
$\gamma_3 = 0.5$	−0.002	0.045	−0.066	0.085	0.006	0.058	0.003	0.044	−0.071	0.090	0.007	0.062
$\sigma_\delta^2 = 1$	−0.025	0.072	−0.163	0.177	−0.013	0.118	−0.020	0.077	−0.175	0.190	−0.012	0.114
$\alpha_r = 0.5$					−0.081	0.257					−0.119	0.272
$\beta_{r1} = -1$					−0.005	0.188					−0.027	0.159
$\beta_{r2} = 0.5$					0.031	0.217					0.009	0.219
$\beta_{r3} = 0.5$					0.011	0.292					0.008	0.263
$\gamma_{r1} = -0.7$					0.090	0.485					0.183	0.494
$\gamma_{r2} = -0.7$					−0.069	0.281					−0.084	0.259
$\phi = -1.2$					−0.121	0.321					−0.129	0.297

Table 4. Out-of-sample prediction accuracy in Part 3 of Simulation 2 with $n = 500$.

	BSOI-Full	BSOI-IN	BSOI-NN	STGP	TV
Mechanism 1 of Simulation 1					
Mean	0.843	0.833	0.839	0.822	0.832
Median	0.844	0.832	0.840	0.823	0.834
SD	0.018	0.022	0.019	0.025	0.021
Remove imaging covariate, $X(\cdot)$					
Mean	0.626	0.619	0.624		
Median	0.624	0.620	0.624		
SD	0.042	0.044	0.042		
Remove one covariate, z_1					
Mean	0.811	0.799	0.809	0.794	0.798
Median	0.812	0.800	0.808	0.797	0.796
SD	0.021	0.026	0.021	0.028	0.023

results for the five models are presented in the first panel of Table 4. As expected, BSOI-Full method with fully observed data exhibits the best prediction accuracy among all methods. The proposed BSOI-NN method also shows a satisfactory performance, whereas the other three methods do not perform very well due to the ignorance of nonignorable nonresponse.

We further examine whether the use of imaging and scalar covariates lead to better prediction accuracy. We first remove the imaging covariate from the datasets and re-evaluate the prediction accuracy of all proposed methods. Without the imaging covariate, both STGP and TV models reduce to a conventional linear regression model and are not considered in this comparison. The obtained results are reported in the second panel of Table 4, showing that the prediction accuracy of all methods drops significantly without the imaging covariate. We then remove one covariate, z_1, from the datasets and re-evaluate the prediction performance of all methods. The obtained results are shown in the third panel of Table 4, depicting a lower prediction accuracy of all methods when ignoring a scalar covariate. The above analyses confirm the power of using imaging and scalar covariates in terms of prediction.

5. The Alzheimer's Disease Neuroimaging Initiative Data

The proposed BSOI-NN method was applied to the ADNI dataset as described in the introduction section. The goal of this analysis is to investigate whether the baseline imaging and scalar covariates can accurately predict the RAVLT learning scores at the 36th month. The learning scores are subject to 45.6% nonresponses. The dataset consists of 802 subjects with 223 NC, 391 MCI patients, and 188 AD patients. Among them, 467 are males (mean age, 75.52 ± 6.78), and 335 are females (mean age, 74.78 ± 6.81).

MRI data were collected from each subject. These MRI images include the standard T1-weighted images obtained through volumetric three-dimensional sagittal magnetization prepared rapid gradient-echo or equivalent protocols with varying resolutions. The MRI images were generated across a variety of 1.5 Tesla MRI scanners with individualized protocols. The following parameters form a typical MRI protocol: repetition time = 2400 ms, inversion time = 1000 ms, flip angle = 8°, field of view = 24 cm with a $256 \times 256 \times 170$ acquisition matrix in the x-, y- and z-dimensions, which yields a voxel size of $1.25 \times 1.26 \times 1.2$ mm^3 (Jack et al. 2008). The MRI data were preprocessed with standard steps as follows: Anterior commissure and posterior commissure corrections were performed on original images (McAuliffe et al. 2001). The images were then resampled to the dimension of 256 × 256 × 256. N2 bias field correction was implemented on the reconstructed images to reduce intensity inhomogeneity (Sled, Zijdenbos, and Evans 1998). A hybrid of the brain surface extractor (Shattuck et al. 2001) and brain extraction tool (Smith 2002) was used to address the problems in each method and ensure skull-stripping accuracy. Another intensity inhomogeneity correction was performed following the skull-stripping procedure. Afterward, the cerebellum was removed from the images according to registration by using a manually labeled cerebellum as a standard template. The deformation field obtained during registration was used to generate the 256 × 256 × 256 regional analysis of the volume in normalized space (RAVENS) images. The RAVENS images for different subjects had a unified shape and size in normalized space, and these intensity-based RAVENS maps preserve the local tissue volumes in original MRIs. The brain regions that were expanded during the normalization step will appear darker than their original counterparts because the same amount of tissue was spread over a larger area, and the regions that were decreased in size will appear proportionally brighter (Goldszal et al. 1998). The intensities of different voxels of a RAVENS map

represent the local volume density at different locations relative to the density at the same locations of the template (Davatzikos et al. 2001). Finally, the RAVENS images were downsampled to $128 \times 128 \times 128$ resolution for final statistical analysis.

Four domains of covariates that were motivated from the factors identified as important in the existing literature were considered in this analysis. The imaging covariates are the generated RAVENS images. The other covariates of interest include demographic variables, that is, gender (z_1: male = 1, female = 0), age (z_2), educational level (z_3), race (z_4: white = 1, other = 0), and whether the subject was ever married (z_5: never married = 1, ever married = 0); a biomarker variable that indicates the risk caused by variations in the APOE gene, such as APOE4 (z_6); and a diagnostic variable, such as whether the individual is diagnosed as having MCI or AD (z_7: MCI or AD = 1, NC = 0). We selected educational level as the instrumental variable. Specifically, the learning ability of an elderly adult is usually strongly correlated with their educational level, while the missingness of the learning abilities may be conditionally uncorrelated with the educational level conditional on the learning ability. To assess whether the educational level is a good candidate for the instrumental variable, we performed an experiment by including the educational level into the exponential tilting model in implementing the BSOI-NN method. The results suggest that educational level is highly correlated with the learning ability and conditionally independent of the missing probability. Moreover, we tried several other variables, including marital status and APOE4, as the instrumental variable. The experiments suggest that the results are not very sensitive to different choices of instrumental variable, and the experiments based on educational level show stable performance. The BSOI regression with $n = 802$ subjects and $Q = 7$ covariates was proposed to conduct the analysis.

For statistical inference, we first applied FPCA on the imaging observations to obtain the corresponding eigenvalues, eigenimages, and eigenscores. Then, we considered the eigenscores as known covariates in the BSOI regression. The prior inputs were specified in the same manner as those in the simulation studies. Several pilot runs from well-separated starting values were then performed, and the results showed that the MCMC algorithm converged within 10,000 iterations. After the burn-in phase of 10,000 iterations, another 10,000 iterations were generated to conduct the posterior inference. For the determination of K, the number of eigenimages retained, we rotated K from 1 to 100 and applied BIC for selection. Figure 5 displays the first 30 BIC values, and those with K over 30 are large and thus are not reported. The BSOI-NN model with $K = 5$ eigenimages exhibits the best fit to the ADNI dataset.

Table 5 presents the estimation results obtained from the BSOI-NN and BSOI-IN models. Several predictors, such as educational level and APOE4, were apparently significant for BSOI-NN but nonsignificant for BSOI-IN. However, these variables were previously recognized as important influential factors for the cognitive ability of elderly people, especially educational level (Lee et al. 2003; Lièvre, Alley, and Crimmins 2008), marital status (Helmer et al. 1999; Petersen et al. 2010), and APOE4 status (Bekris et al. 2010; Petersen et al. 2010). Moreover, ϕ is highly significant in the BSOI-NN model, indicating the strong nonignorability of the missing mechanism. These observations confirm that the missing mechanism in this application is nonignorable, and the analysis can be misleading if the nonignorable missing mechanism is disregarded.

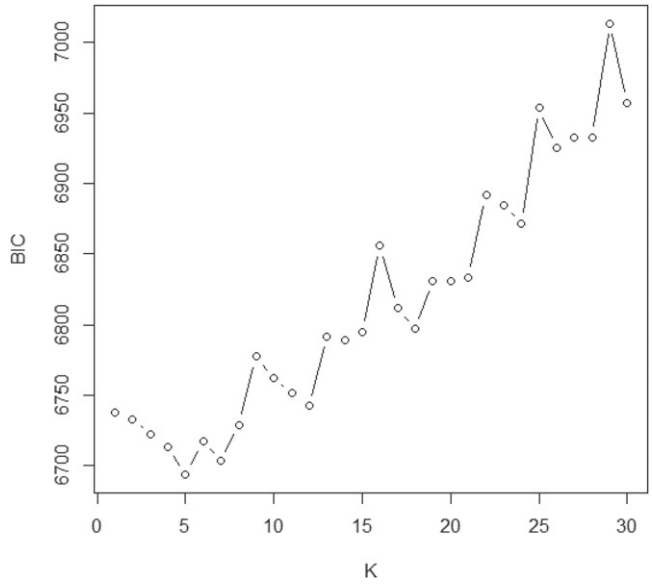

Figure 5. First 30 BIC scores of the BSOI-NN method with different number K of eigenimages in the ADNI data analysis.

Table 5. Bayesian parameter estimates for the BSOI-NN and BSOI-IN models in the analysis of the ADNI dataset.

		BSOI-NN		BSOI-IN	
Covariates	Para	Est	SD	Est	SD
Scalar on image regression					
eigenimage 1	β_1	0.720*	0.158	0.530*	0.137
eigenimage 2	β_2	0.668*	0.152	0.330*	0.132
eigenimage 3	β_3	0.665*	0.140	0.267*	0.122
eigenimage 4	β_4	0.502*	0.151	0.091	0.135
eigenimage 5	β_5	0.870*	0.149	0.434*	0.131
gender	γ_1	−0.190	0.332	−0.173	0.289
age	γ_2	0.066*	0.023	0.031	0.023
educational level	γ_3	0.152*	0.044	0.071	0.043
race	γ_4	−0.035	0.534	−0.127	0.454
whether ever married	γ_5	−2.605*	0.879	−1.168	0.816
APOE4	γ_6	−0.482*	0.202	−0.327	0.182
whether have MCI or AD	γ_7	−3.332*	0.295	−2.291*	0.245
Exponential tilting model					
eigenimage 1	β_{r1}	0.147	0.204		
eigenimage 2	β_{r2}	−0.054	0.176		
eigenimage 3	β_{r3}	−0.207	0.162		
eigenimage 4	β_{r4}	−0.256	0.150		
eigenimage 5	β_{r5}	0.062	0.182		
gender	γ_{r1}	−0.043	0.431		
age	γ_{r2}	0.025	0.017		
race	γ_{r3}	−0.610	0.586		
whether ever married	γ_{r4}	−0.055	0.822		
APOE4	γ_{r5}	−0.051	0.241		
whether have MCI or AD	γ_{r6}	−1.105*	0.510		
learning score	ϕ	−1.287*	0.215		

*Zero is not contained in the 95% credibility interval.

We tested the importance of the instrumental variable by excluding it (i.e., educational level) in the BSOI-NN analysis. Computationally, several MCMC chains did not converge because the instrumental variable was not included, while almost all of the Bayesian parameters estimated became insignificant. Such behavior did not occur for the BSOI-NN model with educational level as the instrumental variable.

Table 6. Bayesian parameter estimates for the BSOI-NN and BSOI-IN models in the analysis of the ADNI dataset without the instrumental variable, educational level.

		BSOI-NN		BSOI-IN	
Covariates	Para	Est	SD	Est	SD
Scalar on image regression					
eigenimage 1	β_1	0.653*	0.155	0.497*	0.129
eigenimage 2	β_2	0.600*	0.171	0.326*	0.124
eigenimage 3	β_3	0.547*	0.160	0.236*	0.115
eigenimage 4	β_4	0.408*	0.180	0.071	0.131
eigenimage 5	β_5	0.767*	0.182	0.407*	0.130
gender	γ_1	−0.050	0.325	−0.132	0.277
age	γ_2	0.047*	0.019	0.023	0.020
educational level	γ_3				
race	γ_4	0.050	0.500	−0.121	0.456
whether ever married	γ_5	−2.556*	0.897	−1.327	0.802
APOE4	γ_6	−0.523	0.198	−0.357	0.174
whether have MCI or AD	γ_7	−3.314*	0.379	−2.357*	0.236
Exponential tilting model					
eigenimage 1	β_{r1}	0.108	0.168		
eigenimage 2	β_{r2}	−0.112	0.179		
eigenimage 3	β_{r3}	−0.245	0.156		
eigenimage 4	β_{r4}	−0.282	0.152		
eigenimage 5	β_{r5}	−0.060	0.208		
gender	γ_{r1}	−0.003	0.373		
age	γ_{r2}	0.012	0.019		
race	γ_{r3}	−0.483	0.581		
whether ever married	γ_{r4}	0.118	0.795		
APOE4	γ_{r5}	−0.051	0.222		
whether have MCI or AD	γ_{r6}	−0.648	0.777		
learning score	ϕ	−1.046*	0.429		

*Zero is not contained in the 95% credibility interval.

Moreover, even for converged MCMC chains, the posterior standard deviation of the tilting parameter ϕ was estimated to be 0.429 under the absence of the instrumental variable. See Table 6 for details. Such a standard deviation value is much larger than that of BSOI-NN with the instrumental variable, indicating that ϕ may be unidentifiable. This result is consistent with the findings in Shao and Wang (2016), revealing the importance of the instrumental variable.

Based on the estimation results, we obtained the following findings. In the exponential tilting model, ϕ is significantly negative, implying that subjects with weak cognitive ability are likely to have future nonresponses. In addition, several elements of scalar covariates and imaging predictors significantly affect the probability of nonresponse. In BSOI regression, the scalar covariates exhibit diverse effects. Education has a positive effect on the learning ability of the elderly people, whereas the status of never married, APOE4, and the clinical outcome of MCI or AD exhibit negative effects. These findings agree with those in the current medical literature (e.g., Helmer et al. 1999; Lee et al. 2003; Bekris et al. 2010). Surprisingly, age has a slightly positive effect on learning ability. One possible explanation is that the ADNI study focused on the population of elderly people, and healthy elderly people tend to possess better cognitive ability and to live longer than unhealthy ones. Nonetheless, further investigation is required to understand this unexpected result.

Regarding the imaging predictors, all five of the retained eigenimages exhibit highly significant effects on the learning scores, indicating that the imaging covariate is indeed an important risk factor for the learning ability of elderly people. $\beta(\cdot)$ was calculated via the Karhumen–Loeve expansion with estimated coefficients β_k and eigenimages (see Figure 6 for the first eigenimage as an example) from FPCA. For clarity, $\beta(\cdot)$ is depicted in Figure 7 such that its positive part is separated from its negative part, which denotes the positive and negative effects of the corresponding brain regions involved in learning ability, respectively. The results are in good agreement with related results reported in the existing literature. For instance, the positive effects of "frontal gyrus" (Figure 8, the top row), "superior temporal gyrus and insula" (Figure 8, the middle row), and "medial temporal lobe, perirhinal and entorhinal cortex" (Figure 8, the bottom row) are relatively large. This finding is consistent with the results in neuroscience studies, which revealed that MCI and AD are negatively associated with the volume or cortical thickness of the frontal brain regions (Hämäläinen et al. 2007; Whitwell et al. 2007; Im et al. 2008b), superior temporal regions (Harasty et al. 1999), insula regions (Foundas et al. 1997), medial temporal lobe (Visser et al. 2002), and perirhinal and entorhinal cortex (Yilmazer-Hanke and Joachim 1999). In brain regions that exhibit negative effects, the "lateral ventricle and caudate nucleus" (Figure 9, the top row), and "central sulcus" (Figure 9, the bottom row) are the most evident. These results are also supported by previous findings, which indicated that the sulcal span (Im et al. 2008a; Liu et al. 2012) and enlargement of the lateral ventricle (McKhann et al. 1984; Feng et al. 2004; Nestor et al. 2008; Ertekin et al. 2016) are associated with the decline in cognitive functions in AD and MCI patients. Notably, the proposed method reveals a significant negative effect of the volume of caudate nucleus on the learning ability of older people, which agrees with the very recent results in cognitive neurology (Persson et al. 2017, 2018).

In addition to the association analysis on the ADNI data, we also evaluated the out-of-sample prediction performance of the proposed method. We randomly partitioned the ADNI dataset into a training subset with $n_1 = 401$ and a testing subset with $n_2 = 401$, ensuring that both subsets exhibit the same nonresponse rate. For BSOI-NN and BSOI-IN methods, we used the training subset with nonresponse to fit model parameters. For STGP and TV methods, we discarded the samples with missing responses in the training subset and applied the remaining observations to fit model parameters. Unlike the simulation study, we did not have the true values of missing responses in the testing subset for validation. As a remedy, for a subject with no learning score at the 36th month, we obtained his/her most recently observed learning score as the true value at the 36th month, such as that from the 30th month, 24th month, or earlier. We repeated the above random partition and analysis for 100 times and computed the mean prediction accuracy (standard deviation) for BSOI-NN, BSOI-IN, STGP, and TV as 0.540 (0.028), 0.533 (0.029), 0.526 (0.030), and 0.527 (0.032), respectively. The results suggest that it is crucial to consider the nonignorable nonresponse in analyzing the ADNI dataset. We also examined whether the medical images are powerful in terms of prediction. We discarded the imaging covariates from the ADNI dataset and repeated the partition procedures and analyses. The prediction accuracy for BSOI-NN and BSOI-IN were computed as 0.508 (0.033) and 0.498 (0.033), respectively, confirming that the medical images are crucial predictors for the learning scores.

When predicting future learning scores, the baseline learning scores are often useful due to within-subject correlation.

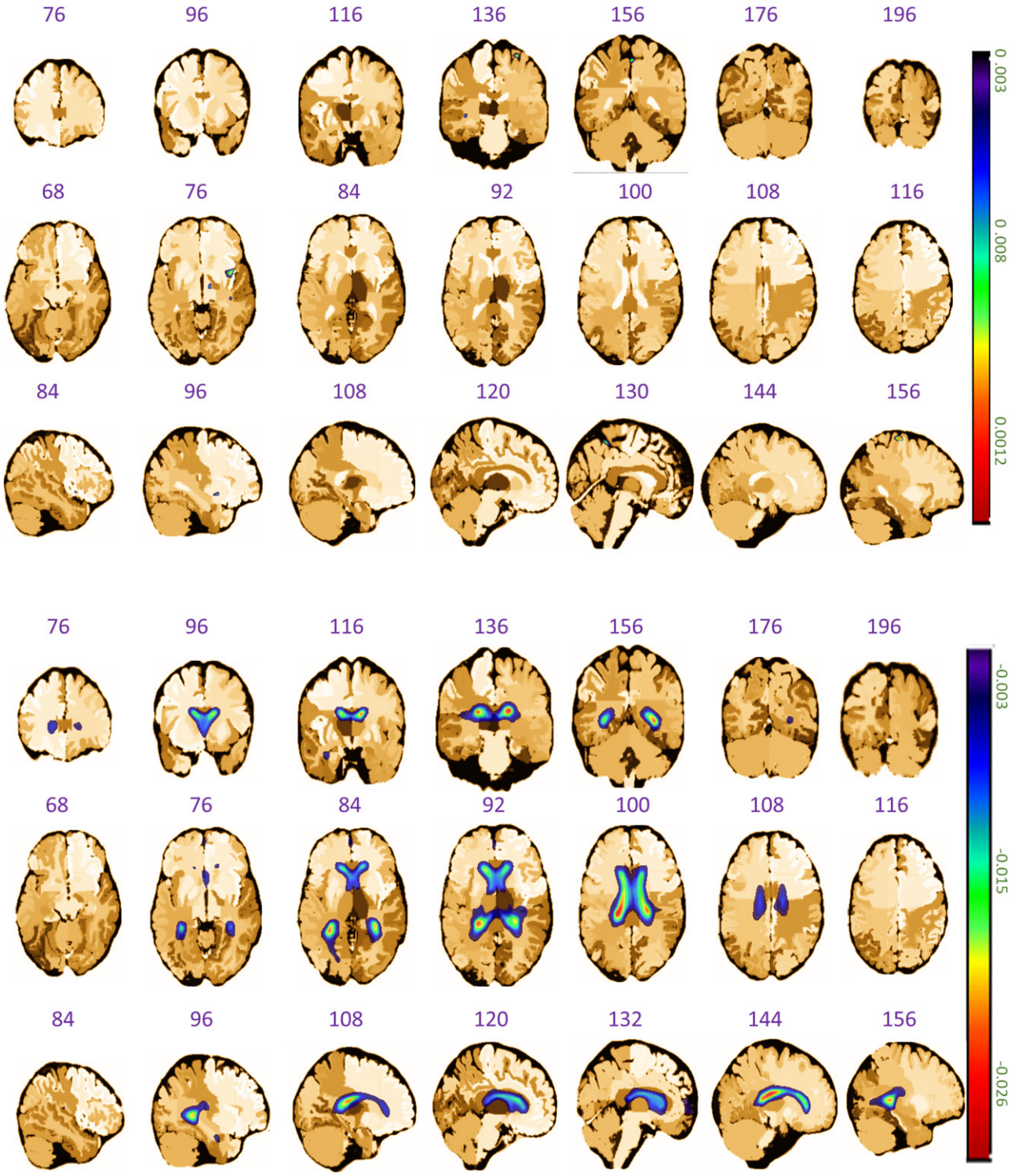

Figure 6. The coronal, axial, and sagittal planes of the estimated first eigenimage in the analysis of ADNI dataset, where the upper three rows represent the positive values and the lower three rows denote the negative values. Note that the sign of an eigenimage is not identifiable in FPCA, and the +/− signs are used here only for separating the regions of eigenimages that exhibit opposite signs.

We therefore further considered the baseline learning scores (z_8) as a covariate in the BSOI model. This setting essentially tests the influence of baseline covariates on the change in the RAVLT learning scores at the 36th month relative to the baseline learning scores. Table 7 presents the estimation results, which are consistent with those of the previous analysis. As expected, elderly people with higher baseline learning scores tend to exhibit better learning ability in the follow-up phase. Notably, although the baseline learning scores exhibit a positive effect on the nonresponse probability, their strong negative mediation effect through the current learning scores on the nonresponse probability implies a negative total effect on the nonresponse probability. To verify this mediation effect, we separately analyzed the exponential tilting model by excluding the

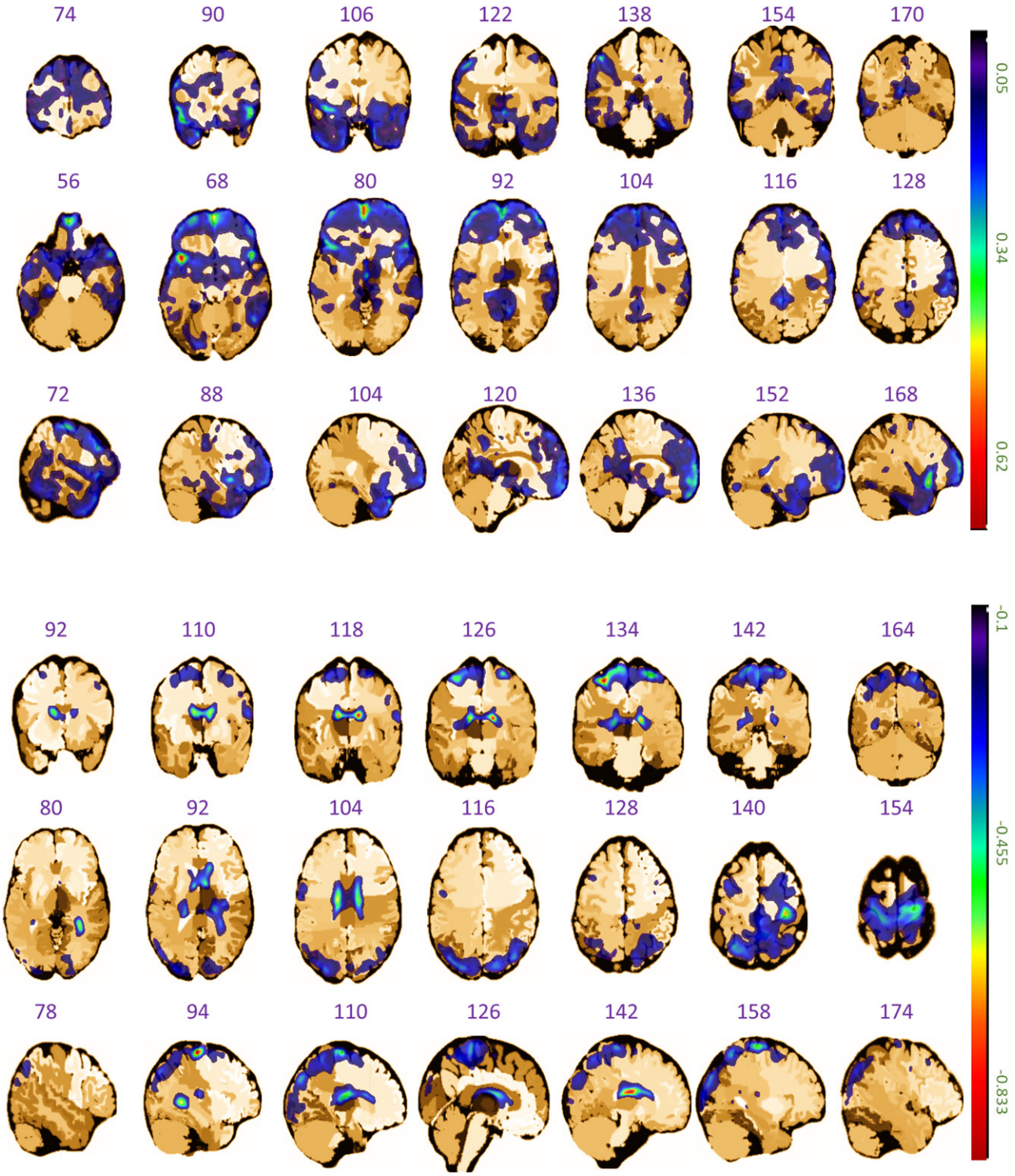

Figure 7. The coronal, axial, and sagittal planes of the estimated coefficient image $\beta(\cdot)$ in the analysis of ADNI dataset, where the upper three rows represent the positive values and the lower three rows denote the negative values.

current learning scores y but retaining the rest. The effect of the baseline learning scores on the nonresponse probability became negative (-0.145^*), which confirms the mediation effect and agrees with the results presented in Figure 1. In terms of out-of-sample prediction, we again repeated the random partition procedures for 100 times and computed the prediction accuracy for BSOI-NN, BSOI-IN, STGP, and TV as 0.615 (0.026), 0.610 (0.025), 0.606 (0.028), and 0.608 (0.031), respectively. These results indicate that baseline learning scores are predictive of future learning scores. It is reasonable to further infer that in addition to the baseline learning scores, the learning scores at other months previous to the 36th month should also be predictive of the learning scores at the 36th month. However, there are considerable missing values for the learning scores at these months, for example, the nonresponse rate at the 24th month is 23.4%. The proposed method can be further improved

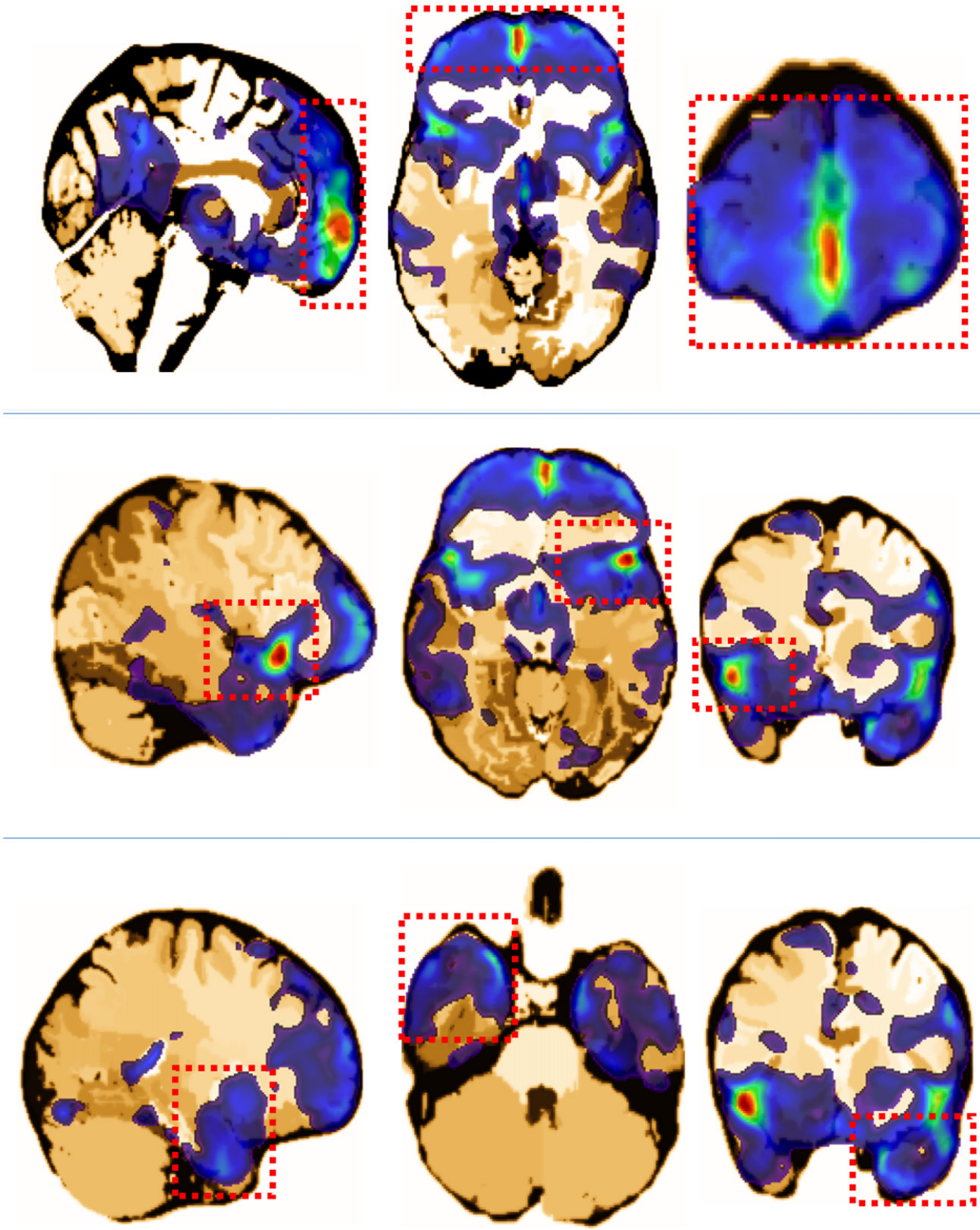

Figure 8. The coronal, axial, and sagittal planes of several positive parts of the estimated coefficient image $\beta(\cdot)$, which are located in the brain regions of the "frontal gyrus" (top row), "superior temporal gyrus and insula" (middle row), and "medial temporal lobe, perirhinal and entorhinal cortex" (bottom row), respectively.

by allowing for nonignorable missing covariates so that the learning scores before the 36th month can be adjusted in the prediction model. We have acknowledged this limitation in the discussion section and included it in our research agenda.

Finally, we evaluated the validity of the conducted analyses through two robustness checks. The first one aims at assessing the impact of the model assumption that $\beta(\cdot)$ can be well represented by the leading eigenimages of $X(\cdot)$. As mentioned in Part 2 of Simulation 2, we may re-estimate $\beta(\cdot)$ under a different assumption. To this end, we used the Bayesian STGP method proposed by Kang, Reich, and Staicu (2018) to estimate $\beta(\cdot)$ in the scalar-on-image regression. The STGP method assumes $\beta(\cdot)$

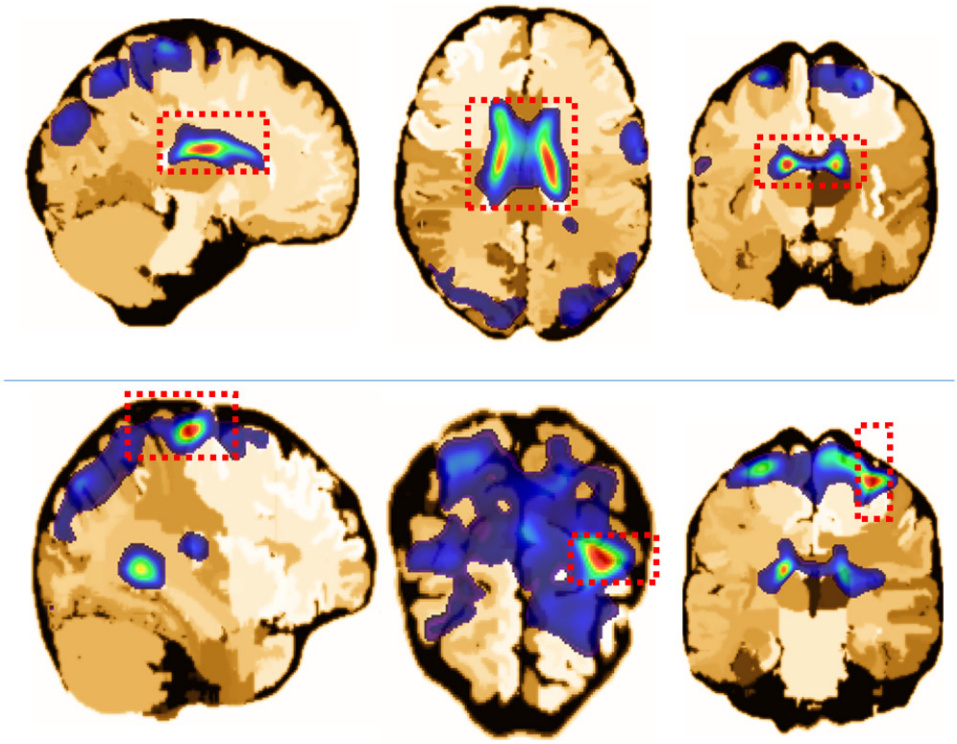

Figure 9. The coronal, axial, and sagittal planes of several negative parts of the estimated coefficient image $\beta(\cdot)$, which are located in the brain regions of "lateral ventricle and caudate nucleus" (top row) and "central sulcus" (bottom row), respectively.

Table 7. Bayesian parameter estimates for the BSOI-NN and BSOI-IN models with the consideration of the baseline learning scores.

		BSOI-NN		BSOI-IN	
Covariates	Para	Est	SD	Est	SD
Scalar on image regression					
eigenimage 1	β_1	0.691*	0.145	0.525*	0.130
eigenimage 2	β_2	0.598*	0.140	0.304*	0.125
eigenimage 3	β_3	0.590*	0.132	0.248*	0.116
eigenimage 4	β_4	0.500*	0.149	0.116	0.125
eigenimage 5	β_5	0.759*	0.145	0.381*	0.124
gender	γ_1	0.026	0.311	0.031	0.270
age	γ_2	0.055*	0.018	0.028	0.021
educational level	γ_3	0.105*	0.036	0.044	0.038
race	γ_4	−0.063	0.466	−0.116	0.423
whether ever married	γ_5	−2.153*	0.826	−0.947	0.785
APOE4	γ_6	−0.267	0.193	−0.155	0.172
whether have MCI or AD	γ_7	−2.294*	0.318	−1.570*	0.253
baseline learning score	γ_8	0.407*	0.050	0.317*	0.045
Exponential tilting model					
eigenimage 1	β_{r1}	0.184	0.205		
eigenimage 2	β_{r2}	−0.084	0.195		
eigenimage 3	β_{r3}	−0.238	0.166		
eigenimage 4	β_{r4}	−0.226	0.188		
eigenimage 5	β_{r5}	0.018	0.192		
gender	γ_{r1}	0.112	0.404		
age	γ_{r2}	0.013	0.017		
race	γ_{r3}	−0.553	0.608		
whether ever married	γ_{r4}	0.180	0.845		
APOE4	γ_{r5}	0.029	0.254		
whether have MCI or AD	γ_{r6}	−0.873	0.591		
baseline learning score	γ_{r7}	0.164*	0.082		
learning score	ϕ	−1.257*	0.225		

*Zero is not contained in the 95% credibility interval.

to be piecewise-smooth, sparse and continuous, and estimates $\beta(\cdot)$ on the voxel level. However, as mentioned above, this voxelwise method is extremely computationally demanding. Therefore, we applied the STGP procedure to estimate $\beta(\cdot)$ only in the scalar-on-image regression and utilized the default settings of the STGP procedure in the R package "STGP." The results based on the STGP method are provided in Appendix D. Two findings are obtained. First, the estimates of the parameters other than $\beta(\cdot)$ are largely consistent with those in Table 6 under the proposed BSOI-NN method. Second, the regions of $\beta(\cdot)$ detected by the STGP method largely overlap with those estimated by the proposed BSOI-NN method. Specifically, 81.3% (2375 among 2922 voxels) of the regions in $\beta(\cdot)$ detected to be positive by the STGP method were also estimated to be positive by the proposed BSOI-NN method, and 69.5% (332 among 478 voxels) of the negative regions detected by the STGP method were also estimated to be negative by the proposed BSOI-NN method. Notably, the estimated $\beta(\cdot)$ using STGP is highly sparse. Therefore, many regions, such as lateral ventricle, caudate nucleus, and central sulcus, which have been detected by our method and well validated by the existing literature (McKhann et al. 1984; Feng et al. 2004; Im et al. 2008a; Nestor et al. 2008; Liu et al. 2012; Ertekin et al. 2016), were not detected by STGP. These results indicate that the model assumption on $\beta(\cdot)$ may be plausible and the proposed BSOI model framework is potentially useful for substantive studies in the presence of ultrahigh dimensional imaging data and nonignorable missingness.

The second robustness check excluded the diagnostic status (z_7) from both the main analysis and the analysis with the baseline learning scores. The baseline diagnostic status is partially determined by the baseline RAVLT scores, so it is strongly correlated with the baseline and future RAVLT scores. Thus, the diagnostic status could be a substantial confounding factor in the analysis and may significantly affect the results. The estimation results are consistent with those of the above

analyses, and thus are not reported to save space. This may provide further evidence of the validity of the previous analyses.

6. Discussion

We develop a Bayesian method for analyzing BSOI regression with nonignorable nonresponses. Ultrahigh-dimensional imaging data are considered in the regression model to explain clinical outcomes of interest. We propose an exponential tilting model with scalar and imaging predictors to examine their effects on the probability of nonresponse. An instrumental variable is introduced to the missing data model to aid model identifiability and estimation. FPCA for high-dimensional imaging observations is employed to lower the dimensionality of the imaging regression and missing data model to simplify the model representation. An efficient Bayesian method is developed and used with MCMC techniques to conduct statistical inference. The effectiveness of the proposed method is demonstrated through simulation studies. We believe that its application to the ADNI dataset provides new insights into AD risk factors.

The present study has limitations. First, the FPCA method used to reduce model dimensionality relies on the presmoothing of the covariance operator, while the spatial structures of the brain images are disregarded. The combination of Ising and Markov random field priors developed by Goldsmith, Huang, and Crainiceanu (2014) may be applied to accommodate the spatial correlations. However, such priors introduce high computational challenges under ultrahigh-dimensional settings. The feasibility of such development requires further investigation. Second, we assume $\beta(\cdot)$ to be well presented by the eigenimages of $X(\cdot)$. This may not be the case in certain circumstances. In fact, $\beta(\cdot)$ may have unknown jumps and edges. We can consider a generalization of the current approach by applying functional penalization methods to directly estimate $\beta(\cdot)$ (e.g., Crambes, Kneip, and Sarda 2009; Wang and Zhu 2017; Kang, Reich, and Staicu 2018). Third, to better explain the current learning scores, the proposed methodology can be further improved by allowing for nonignorable missing covariates so that the historical learning scores that are subjected to nonignorable missingness can be considered in the regression model. Fourth, longitudinal nonignorable nonresponse has been recognized as an important research problem in the literature (see, e.g., Diggle and Kenward 1994). We may extend the proposed BSOI-NN framework to handle longitudinal nonignorable nonresponse so that the entire trajectory of the learning scores that are subjected to nonignorable missingness can be linked with imaging covariates. Finally, we consider a parametric setting in the BSOI regression and missing mechanism. We may extend the proposed framework to a more sophisticated semiparametric context. Substantial efforts are required for these developments.

Appendix A: Proof for Theorem 1

Suppose that the following two equations hold for all $(y, X(v), \mathbf{z}^*) \in S$ and $(u_1, u_2) : u_1 \neq u_2$.

$$\begin{aligned}
&\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r1})p(y|X(\cdot),\mathbf{z}^*,u_1;\boldsymbol{\theta}_{y1})\\
&=\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r2})p(y|X(\cdot),\mathbf{z}^*,u_1;\boldsymbol{\theta}_{y2}),\\
&\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r2})p(y|X(\cdot),\mathbf{z}^*,u_2;\boldsymbol{\theta}_{y2})\\
&=\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r1})p(y|X(\cdot),\mathbf{z}^*,u_2;\boldsymbol{\theta}_{y1}).
\end{aligned} \tag{A.1}$$

Multiplying the two equations gives

$$\begin{aligned}
&\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r1})p(y|X(\cdot),\mathbf{z}^*,u_1;\boldsymbol{\theta}_{y1})\\
&\Pr(r=0|y,X(v),\mathbf{z}^*;\boldsymbol{\theta}_{r2})p(y|X(\cdot),\mathbf{z}^*,u_2;\boldsymbol{\theta}_{y2})\\
&=\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r2})p(y|X(\cdot),\mathbf{z}^*,u_1;\boldsymbol{\theta}_{y2})\\
&\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r1})p(y|X(\cdot),\mathbf{z}^*,u_2;\boldsymbol{\theta}_{y1}).
\end{aligned} \tag{A.2}$$

Together with condition (C1) of Theorem 1, it follows that

$$\begin{aligned}
&p(y|X(\cdot),\mathbf{z}^*,u_1;\boldsymbol{\theta}_{y1})p(y|X(\cdot),\mathbf{z}^*,u_2;\boldsymbol{\theta}_{y2})\\
&=p(y|X(\cdot),\mathbf{z}^*,u_1;\boldsymbol{\theta}_{y2})p(y|X(\cdot),\mathbf{z}^*,u_2;\boldsymbol{\theta}_{y1})
\end{aligned} \tag{A.3}$$

holds for all $(y, X(\cdot), \mathbf{z}^*)$. Together with condition (C2), we have $\boldsymbol{\theta}_{y1} = \boldsymbol{\theta}_{y2}$. Then, we obtain from (A.1) that

$$\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r1}) = \Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r2})$$

for all $(y, X(\cdot), \mathbf{z}^*)$. Together with condition (C3), we have $\boldsymbol{\theta}_{r1} = \boldsymbol{\theta}_{r2}$, and the identifiability is obtained.

Appendix B: Proof for Proposition 1

(I) Condition (C1) holds, because for all $(\boldsymbol{\theta}_r, X(\cdot), \mathbf{z}^*, y)$ we have

$$\begin{aligned}
&\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_r)\\
&=\frac{1}{1+\exp(\alpha_r+\int_{\mathcal{V}}X(v)\beta_r(v)dv+\boldsymbol{\gamma}_r^T\mathbf{z}^*+\phi_1 y)}>0.
\end{aligned}$$

(II) The fact that

$$\Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r1}) = \Pr(r=0|y,X(\cdot),\mathbf{z}^*;\boldsymbol{\theta}_{r2}),$$

for all$(y, X(\cdot), \mathbf{z}^*)$ is equivalent to

$$\begin{aligned}
&\alpha_{r1}+\int_{\mathcal{V}}X(v)\beta_{r1}(v)dv+\boldsymbol{\gamma}_{r1}^T\mathbf{z}^*+\phi_1 y\\
&=\alpha_{r2}+\int_{\mathcal{V}}X(v)\beta_{r2}(v)dv+\boldsymbol{\gamma}_{r2}^T\mathbf{z}^*+\phi_2 y,
\end{aligned}$$

for all $(y, X(\cdot), \mathbf{z}^*) \Longrightarrow$

$$\begin{aligned}
&(\alpha_{r1}-\alpha_{r2})+\int_{\mathcal{V}}X_1(v)\big(\beta_{r1}(v)-\beta_{r2}(v)\big)dv\\
&\qquad+(\boldsymbol{\gamma}_{r1}-\boldsymbol{\gamma}_{r2})^T\mathbf{z}_1^*+(\phi_1-\phi_2)y_1=0,
\end{aligned}$$

$$\begin{aligned}
&(\alpha_{r1}-\alpha_{r2})+\int_{\mathcal{V}}X_1(v)\big(\beta_{r1}(v)-\beta_{r2}(v)\big)dv\\
&\qquad+(\boldsymbol{\gamma}_{r1}-\boldsymbol{\gamma}_{r2})^T\mathbf{z}_1^*+(\phi_1-\phi_2)y_2=0,
\end{aligned}$$

$$\begin{aligned}
&(\alpha_{r1}-\alpha_{r2})+\int_{\mathcal{V}}X_2(v)\big(\beta_{r1}(v)-\beta_{r2}(v)\big)dv\\
&\qquad+(\boldsymbol{\gamma}_{r1}-\boldsymbol{\gamma}_{r2})^T\mathbf{z}_1^*+(\phi_1-\phi_2)y_1=0,
\end{aligned}$$

$$\begin{aligned}
&(\alpha_{r1}-\alpha_{r2})+\int_{\mathcal{V}}X_1(v)\big(\beta_{r1}(v)-\beta_{r2}(v)\big)dv\\
&\qquad+(\boldsymbol{\gamma}_{r1}-\boldsymbol{\gamma}_{r2})^T\mathbf{z}_2^*+(\phi_1-\phi_2)y_1=0,
\end{aligned}$$

for all $(y_1, y_2, X_1(\cdot), X_2(\cdot), \mathbf{z}_1^*, \mathbf{z}_2^*) \Longrightarrow$

$$(\phi_1 - \phi_2)(y_1 - y_2) = 0,$$

$$(\boldsymbol{\gamma}_{r1} - \boldsymbol{\gamma}_{r2})^T(\mathbf{z}_1^* - \mathbf{z}_2^*) = 0,$$

$$\int_{\mathcal{V}} \big(X_1(v) - X_2(v)\big)\big(\beta_{r1}(v) - \beta_{r2}(v)\big)dv = 0,$$

for all $(y_1, y_2, X_1(\cdot), X_2(\cdot), \mathbf{z}_1^*, \mathbf{z}_2^*)$. It follows that $\phi_1 - \phi_2 = 0$ because $y_1 \neq y_2$, and

$$\boldsymbol{\gamma}_{r1} - \boldsymbol{\gamma}_{r2} \in \mathcal{L}(\mathcal{Z}^*, \mathcal{D}(\boldsymbol{\gamma}_r); \mathbb{R}^{Q-1})^{\perp},$$
$$\beta_{r1}(\cdot) - \beta_{r2}(\cdot) \in \mathcal{L}(\mathcal{X}, \mathcal{D}(\beta_r(\cdot)); L_2(\mathcal{V}))^{\perp}.$$

Given $\mathcal{L}(\mathcal{X}, \mathcal{D}(\beta_r(\cdot)); L_2(\mathcal{V}))^{\perp}$ and $\mathcal{L}(\mathcal{Z}^*, \mathcal{D}(\boldsymbol{\gamma}_r); \mathbb{R}^{Q-1})^{\perp}$ are both zero, we have $\phi_1 = \phi_2, \boldsymbol{\gamma}_{r1} = \boldsymbol{\gamma}_{r2}$, and $\beta_{r1}(\cdot) = \beta_{r2}(\cdot)$. Finally, we have $\alpha_{r1} - \alpha_{r2} = 0$.

(III) From the normal distribution assumption of random noise δ_i, the fact that

$$p(y|X(\cdot), \mathbf{z}^*, u_1; \boldsymbol{\theta}_{y1})p(y|X(\cdot), \mathbf{z}^*, u_2; \boldsymbol{\theta}_{y2})$$
$$\equiv p(y|X(\cdot), \mathbf{z}^*, u_1; \boldsymbol{\theta}_{y2})p(y|X(\cdot), \mathbf{z}^*, u_2; \boldsymbol{\theta}_{y1}),$$

for all $(y, X(\cdot), \mathbf{z}^*, u_1, u_2)$ is equivalent to

$$\frac{\big(y - \alpha_1 - \int_{\mathcal{V}} X(v)\beta_1(v)dv - \gamma_{*1}^T\mathbf{z}^* - \gamma_{u1}^T u_1\big)^2}{-\big(y - \alpha_1 - \int_{\mathcal{V}} X(v)\beta_1(v)dv - \gamma_{*1}^T\mathbf{z}^* - \gamma_{u1}^T u_2\big)^2}{2\sigma_1}$$
$$= \frac{\big(y - \alpha_2 - \int_{\mathcal{V}} X(v)\beta_2(v)dv - \gamma_{*2}^T\mathbf{z}^* - \gamma_{u2}^T u_1\big)^2}{-\big(y - \alpha_2 - \int_{\mathcal{V}} X(v)\beta_2(v)dv - \gamma_{*2}^T\mathbf{z}^* - \gamma_{u2}^T u_2\big)^2}{2\sigma_2},$$

for all $(y, X(\cdot), \mathbf{z}^*, u_1, u_2) \Longrightarrow$

$$\frac{1}{\sigma_1}\Big[\gamma_{u1}\big(\gamma_{u1}u_1 + \gamma_{u1}u_2 - 2y + 2\alpha_1 + 2\int_{\mathcal{V}} X(v)\beta_1(v)dv + 2\gamma_{*1}\mathbf{z}^*\big)\Big]$$
$$= \frac{1}{\sigma_2}\Big[\gamma_{u2}\big(\gamma_{u2}u_1 + \gamma_{u2}u_2 - 2y + 2\alpha_2 + 2\int_{\mathcal{V}} X(v)\beta_2(v)dv + 2\gamma_{*2}\mathbf{z}^*\big)\Big],$$

for all $(y, X(\cdot), \mathbf{z}^*, u_1, u_2)$. Similar to the proof of (II), we have $\gamma_{u1}/\sigma_1 = \gamma_{u2}/\sigma_2, \beta_1(\cdot) = \beta_2(\cdot), \gamma_{*1} = \gamma_{*2}, \gamma_{u1} = \gamma_{u2}, \alpha_1 = \alpha_2$ using the fact that $|\gamma_{u1}|/\sigma_1 \geq \epsilon/\sigma_1 > 0$, which completes the proof.

Note that in this proposition the scalar u can be generalized to multidimensional in the model setting. Proof will be similar if we choose $\mathbf{u}_1 = u_1\mathbf{e}_j, \mathbf{u}_2 = u_2\mathbf{e}_j$ to replace u_1, u_2 in the above proof, where $\mathbf{e}_j, j = 1, 2, \ldots, d_u$, are the basis of the space $\mathcal{U}$.

Appendix C: Full Conditional Distributions

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T, \boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^T$, and a vector $\mathbf{a}_{(-k)}$ denote a sub-vector of $\mathbf{a}$ excluding the kth element.

(I) The conditional distributions of unknown parameters in $\boldsymbol{\theta}_y$ are given as follows:

$$p(\alpha|\cdot) \overset{D}{=} \mathrm{N}\left(\Sigma_\alpha[\alpha_0/\sigma_{\alpha 0}^2 + \sum_{i=1}^n (y_i - \boldsymbol{\gamma}^T\mathbf{z}_i - \boldsymbol{\beta}^T\boldsymbol{\xi}_i)/\sigma^2], \Sigma_\alpha\right),$$

$$p(\beta_k|\cdot) \overset{D}{=} \mathrm{N}\Big(\Sigma_{\beta_k}[\beta_{k0}/\sigma_{\beta_k 0}^2 + \sum_{i=1}^n (y_i - \alpha - \boldsymbol{\gamma}^T\mathbf{z}_i - \boldsymbol{\beta}_{(-k)}^T\boldsymbol{\xi}_{i(-k)})\xi_{ik}/\sigma^2], \Sigma_{\beta_k}\Big), \quad k = 1, \ldots, K,$$

$$p(\gamma_q|\cdot) \overset{D}{=} \mathrm{N}\Big(\Sigma_{\gamma_q}[\gamma_{q0}/\sigma_{\gamma_q 0}^2 + \sum_{i=1}^n (y_i - \alpha - \boldsymbol{\gamma}_{(-q)}^T\mathbf{z}_{i(-q)} - \boldsymbol{\beta}^T\boldsymbol{\xi}_i)z_{iq}/\sigma^2], \Sigma_{\gamma_q}\Big), \quad q = 1, \ldots, Q,$$

$$p(\sigma^{-2}|\cdot) \overset{D}{=} \mathrm{Gamma}(a_{\sigma 0} + n/2, b_\sigma),$$

where $\Sigma_\alpha = (n/\sigma^2 + 1/\sigma_{\alpha 0}^2)^{-1}$, $\Sigma_{\beta_k} = (\sum_{i=1}^n \xi_{ik}^2/\sigma^2 + 1/\sigma_{\beta_k 0}^2)^{-1}$, $\Sigma_{\gamma_q} = (\sum_{i=1}^n z_{iq}^2/\sigma^2 + 1/\sigma_{\gamma_q 0}^2)^{-1}$, and $b_\sigma = b_{\sigma 0} + \sum_{i=1}^n (y_i - \alpha - \boldsymbol{\gamma}^T\mathbf{z}_i - \boldsymbol{\beta}^T\boldsymbol{\xi}_i)^2/2$.

(II) The conditional distribution of $y_{\mathrm{mis},i}$ is

$$p(y_{\mathrm{mis},i}|\cdot) \propto \frac{\exp\big\{-(y_{\mathrm{mis}} - \alpha - \boldsymbol{\gamma}^T\mathbf{z}_{\mathrm{mis},i} - \boldsymbol{\beta}^T\boldsymbol{\xi}_{\mathrm{mis},i})^2/2\sigma_\delta^2 + r_i(\alpha_r + \boldsymbol{\gamma}_r^T\mathbf{z}_{\mathrm{mis},i}^* + \boldsymbol{\beta}_r^T\boldsymbol{\xi}_{\mathrm{mis},i} + \phi y_{\mathrm{mis},i})\big\}}{1 + \exp(\alpha_r + \boldsymbol{\gamma}_r^T\mathbf{z}_{\mathrm{mis},i}^* + \boldsymbol{\beta}_r^T\boldsymbol{\xi}_{\mathrm{mis},i} + \phi y_{\mathrm{mis},i})},$$

where $\mathbf{z}_{\mathrm{mis},i}^*$ is the covariate vector excluding the instrumental variable and corresponding to the subject with the missing response.

The above conditional distribution is not in a closed form. We propose the use of the Metropolis–Hastings algorithm to sample from it. At the jth iteration with a current value $y_{\mathrm{mis},i}^{(j)}$, a new candidate value $y_{\mathrm{mis},i}$ is generated from the proposed distribution $\mathrm{N}(y_{\mathrm{mis},i}^{(j)}, \sigma_{y_{\mathrm{mis}}}^2\Sigma_{y_{\mathrm{mis}},i})$, where

$$\Sigma_{y_{\mathrm{mis}},i}^{-1} = \sigma^{-2} + \frac{\exp(\alpha_r + \boldsymbol{\gamma}_r^T\mathbf{z}_i^* + \boldsymbol{\beta}_r^T\boldsymbol{\xi}_i)}{(1 + \exp(\alpha_r + \boldsymbol{\gamma}_r^T\mathbf{z}_i^* + \boldsymbol{\beta}_r^T\boldsymbol{\xi}_i))^2}\phi^2.$$

The new candidate value is accepted according to the following probability:

$$\min\left[1, \frac{p(y_{\mathrm{mis},i}|\cdot)}{p(y_{\mathrm{mis},i}^{(j)}|\cdot)}\right].$$

The variance of $\sigma_{y_{\mathrm{mis}}}^2$ is chosen such that the average acceptance rate is approximately 0.25 (Gelman, Roberts, and Gilks 1996).

(III) The conditional distribution of the unknown parameter vector $\boldsymbol{\theta}_r$ is given as follows:

$$p(\boldsymbol{\theta}_r) \propto \frac{\exp\big\{\sum_{i=1}^n r_i(\alpha_r + \boldsymbol{\gamma}_r^T\mathbf{z}_{\mathrm{mis},i}^* + \boldsymbol{\beta}_r^T\boldsymbol{\xi}_{\mathrm{mis},i} + \phi y_{\mathrm{mis},i}) - \frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r0})^T\boldsymbol{\Sigma}_{r0}^{-1}(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0)\big\}}{\prod_{i=1}^n \big\{1 + \exp(\alpha_r + \boldsymbol{\gamma}_r^T\mathbf{z}_{\mathrm{mis},i}^* + \boldsymbol{\beta}_r^T\boldsymbol{\xi}_{\mathrm{mis},i} + \phi y_{\mathrm{mis},i})\big\}}.$$

This distribution is also not in a closed form. We again use the Metropolis–Hastings algorithm to sample from it. The proposed distribution is chosen as $\mathrm{N}(\mathbf{0}, \sigma_r^2\boldsymbol{\Sigma}_r)$, where

$$\boldsymbol{\Sigma}_r^{-1} = \boldsymbol{\Sigma}_{r0}^{-1} + \frac{1}{4}\sum_{i=1}^n \mathbf{e}_i\mathbf{e}_i^T,$$

in which $\mathbf{e}_i = (1, \mathbf{z}_{\mathrm{mis},i}^{*\top}, \boldsymbol{\xi}_{\mathrm{mis},i}^T)^T$, and σ_r^2 is chosen to appropriately control the acceptance rate.

Appendix D: Results of Robustness Check 1

Table D1. Bayesian parameter estimates of the BSOI-NN model with the STGP technique for estimating $\beta(\cdot)$ in the analysis of the ADNI dataset.

		BSOI-NN-STGP	
Covariates	Para	Est	SD
Scalar on image regression			
Imaging covariate	$\beta(\cdot)$	Sparse	
Gender	γ_1	−0.253	0.266
Age	γ_2	0.017	0.019
Educational level	γ_3	0.109*	0.044
Race	γ_4	0.024	0.479
Whether ever married	γ_5	−1.554*	0.728
APOE4	γ_6	−0.663*	0.207
Whether have MCI or AD	γ_7	−3.490*	0.461
Exponential tilting model			
Eigenimage 1	β_{r1}	−0.039	.318
Eigenimage 2	β_{r2}	−0.002	0.018
Eigenimage 3	β_{r3}	−0.374	0.494
Eigenimage 4	β_{r4}	0.547	0.805
Eigenimage 5	β_{r5}	−0.060	0.214
Gender	γ_{r1}	−0.278	0.685
Age	γ_{r2}	−0.105	0.152
Race	γ_{r3}	−0.273	0.173
Whether ever married	γ_{r4}	−0.402*	0.126
APOE4	γ_{r5}	−0.475*	0.135
Whether have MCI or AD	γ_{r6}	−0.357*	0.137
Learning score	ϕ	−0.891*	0.306

*Zero is not contained in the 95% credibility interval.

Funding

Dr. Feng's work was partially supported by NSFC grants 71802166 and 71490722. Dr. Song's work was partially supported by GRF grants 14303017 and 14301918 from the Research Grant Council of the Hong Kong Special Administration Region. Dr. Zhu's work was partially supported by NIH grant MH086633, a grant from the Cancer Prevention Research Institute of Texas, and the endowed Bao-Shan Jing Professorship in Diagnostic Imaging. Dr. Li's work was partially supported by NIH grant MH116527. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or any other funding agency.

References

Allen, G. I., Amoroso, N., Anghel, C., Balagurusamy, V., Bare, C. J., Beaton, D., Bellotti, R., Bennett, D. A., Boehme, K. L., Boutros, P. C., Caberlotto, L., Caloian, C., Campbell, F., Chaibub Neto, E., Chang, Y.-C., Chen, B., Chen, C.-Y., Chien, T.-Y., Clark, T., Das, S., Davatzikos, C., Deng, J., Dillenberger, D., Dobson, R. J. B., Dong, Q., Doshi, J., Duma, D., Errico, R., Erus, G., Everett, E., Fardo, D. W., Friend, S. H., Fröhlich, H., Gan, J., St George-Hyslop, P., Ghosh, S. S., Glaab, E., Green, R. C., Guan, Y., Hong, M.-Y., Huang, C., Hwang, J., Ibrahim, J., Inglese, P., Iyappan, A., Jiang, Q., Katsumata, Y., Kauwe, J. S. K., Klein, A., Kong, D., Krause, R., Lalonde, E., Lauria, M., Lee, E., Lin, X., Liu, Z., Livingstone, J., Logsdon, B. A., Lovestone, S., Ma, T.-W., Malhotra, A., Mangravite, L. M., Maxwell, T. J., Merrill, E., Nagorski, J., Namasivayam, A., Narayan, M., Naz, M., Newhouse, S. J., Norman, T. C., Nurtdinov, R. N., Oyang, Y.-J., Pawitan, Y., Peng, S., Peters, M. A., Piccolo, S. R., Praveen, P., Priami, C., Sabelnykova, V. Y., Senger, P., Shen, X., Simmons, A., Sotiras, A., Stolovitzky, G., Tangaro, S., Tateo, A., Tung, Y.-A., Tustison, N. J., Varol, E., Vradenburg, G., Weiner, M. W., Xiao, G., Xie, L., Xie, Y., Xu, J., Yang, H., Zhan, X., Zhou, Y., Zhu, F., Zhu, H., Zhu, S., and Alzheimer's Disease Neuroimaging Initiative (2016), "Crowdsourced Estimation of Cognitive Decline and Resilience in Alzheimer's Disease," *Alzheimer's & Dementia*, 12, 645–653. [1574]

Baker, S. G., and Laird, N. M. (1988), "Regression Analysis for Categorical Variables With Outcome Subject to Nonignorable Nonresponse," *Journal of the American Statistical Association*, 83, 62–69. [1575]

Bekris, L. M., Yu, C.-E., Bird, T. D., and Tsuang, D. W. (2010), "Genetics of Alzheimer Disease," *Journal of Geriatric Psychiatry and Neurology*, 23, 213–227. [1587,1588]

Chen, K. (2001), "Parametric Models for Response-Biased Sampling," *Journal of the Royal Statistical Society*, Series B, 63, 775–789. [1575]

Chen, L.-H., and Jiang, C.-R. (2017), "Multi-Dimensional Functional Principal Component Analysis," *Statistics and Computing*, 27, 1181–1192. [1577]

Chiou, J.-M., Zhang, Y.-C., Chen, W.-H., and Chang, C.-W. (2014), "A Functional Data Approach to Missing Value Imputation and Outlier Detection for Traffic Flow Data," *Transportmetrica B: Transport Dynamics*, 2, 106–129. [1576]

Crambes, C., Kneip, A., and Sarda, P. (2009), "Smoothing Splines Estimators for Functional Linear Regression," *The Annals of Statistics*, 37, 35–72. [1593]

Davatzikos, C., Genc, A., Xu, D., and Resnick, S. M. (2001), "Voxel-Based Morphometry Using the Ravens Maps: Methods and Validation Using Simulated Longitudinal Atrophy," *NeuroImage*, 14, 1361–1369. [1587]

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, Series B, 41, 1–15. [1577]

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009), "Multilevel Functional Principal Component Analysis," *The Annals of Applied Statistics*, 3, 458–488. [1576]

Diggle, P., and Kenward, M. G. (1994), "Informative Drop-Out in Longitudinal Data Analysis," *Journal of the Royal Statistical Society*, Series C, 43, 49–93. [1575,1593]

Eddelbuettel, D., and Sanderson, C. (2014), "RcppArmadillo: Accelerating R With High-Performance C++ Linear Algebra," *Computational Statistics & Data Analysis*, 71, 1054–1063. [1580]

Ertekin, T., Acer, N., Köseoğlu, E., Zararsız, G., Sönmez, A., Gümüş, K., and Kurtoğlu, E. (2016), "Total Intracranial and Lateral Ventricle Volumes Measurement in Alzheimer's Disease: A Methodological Study," *Journal of Clinical Neuroscience*, 34, 133–139. [1588,1592]

Feng, R., Wang, H., Wang, J., Shrom, D., Zeng, X., and Tsien, J. Z. (2004), "Forebrain Degeneration and Ventricle Enlargement Caused by Double Knockout of Alzheimer's Presenilin-1 and Presenilin-2," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 8162–8167. [1588,1592]

Ferraty, F., Sued, M., and Vieu, P. (2013), "Mean Estimation With Data Missing at Random for Functional Covariables," *Statistics*, 47, 688–706. [1576]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Methods, Theory, Applications and Implementation*, New York: Springer. [1575]

Foundas, A. L., Leonard, C. M., Mahoney, S. M., Agee, O. F., and Heilman, K. M. (1997), "Atrophy of the Hippocampus, Parietal Cortex, and Insula in Alzheimer's Disease: A Volumetric Magnetic Resonance Imaging Study," *Cognitive and Behavioral Neurology*, 10, 81–89. [1588]

Fraiman, R., Gimenez, Y., and Svarc, M. (2016), "Feature Selection for Functional Data," *Journal of Multivariate Analysis*, 146, 191–208. [1575]

Gelfand, A. E., and Sahu, S. K. (1999), "Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models," *Journal of the American Statistical Association*, 94, 247–253. [1577]

Gelfand, A. E., and Smith, A. F. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [1577]

Gelman, A., Roberts, G., and Gilks, W. (1996), "Efficient Metropolis Jumping Hules," *Bayesian Statistics*, 5, 599–608. [1594]

Gertheiss, J., Goldsmith, J., Crainiceanu, C., and Greven, S. (2013), "Longitudinal Scalar-on-Functions Regression With Application to Tractography Data," *Biostatistics*, 14, 447–461. [1576]

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2015), "Radiomics: Images Are More Than Pictures, They Are Data," *Radiology*, 278, 563–577. [1575]

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011), "Penalized Functional Regression," *Journal of Computational and Graphical Statistics*, 20, 830–851. [1576]

Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2014), "Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection," *Journal of Computational and Graphical Statistics*, 23, 46–64. [1593]

Goldszal, A. F., Davatzikos, C., Pham, D. L., Yan, M. X., Bryan, R. N., and Resnick S. M. (1998), "An Image-Processing System for Qualitative and Quantitative Volumetric Analysis of Brain Images," *Imaging Journal of Computer Assisted Tomography*, 22, 827–837. [1586]

Golub, G. H., and Loan, V. (1996), *Matrix Computations* (Vol. 1), Baltimore, MD: Johns Hopkins University Press. [1577]

Hall, P., and Horowitz, J. L. (2007), "Methodology and Convergence Rates for Functional Linear Regression," *The Annals of Statistics*, 35, 70–91. [1575]

Hämäläinen, A., Tervo, S., Grau-Olivares, M., Niskanen, E., Pennanen, C., Huuskonen, J., Kivipelto, M., Hänninen, T., Tapiola, M., Vanhanen, M., and Hallikainen, M. (2007), "Voxel-Based Morphometry to Detect Brain Atrophy in Progressive Mild Cognitive Impairment," *Neuroimage*, 37, 1122–1131. [1588]

Happ, C., Greven, S., and Schmid, V. J. (2018), "The Impact of Model Assumptions in Scalar-on-Image Regression," *Statistics in Medicine*, 37, 4298–4317. [1583]

Harasty, J. A., Halliday, G. M., Kril, J. J., and Code, C. (1999), "Specific Temporoparietal Gyral Atrophy Reflects the Pattern of Language Dissolution in Alzheimer's Disease," *Brain*, 122, 675–686. [1588]

Helmer, C., Damon, D., Letenneur, L., Fabrigoule, C., Barberger-Gateau, P., Lafont, S., Fuhrer, R., Antonucci, T., Commenges, D., Orgogozo, J., and Dartigues, J. F. (1999), "Marital Status and Risk of Alzheimer's Disease: A French Population-Based Cohort Study," *Neurology*, 53, 1953–1953. [1587,1588]

Horváth, L., and Kokoszka, P. (2012), *Inference for Functional Data With Applications* (Vol. 200), New York: Springer Science. [1575]

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001), "Missing Responses in Generalised Linear Mixed Models When the Missing Data Mechanism Is Nonignorable," *Biometrika*, 88, 551–564. [1575,1577]

——— (2002), "Bayesian Methods for Generalized Linear Models With Covariates Missing at Random," *Canadian Journal of Statistics*, 30, 55–78. [1576]

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005), "Missing-Data Methods for Generalized Linear Models," *Journal of the American Statistical Association*, 100, 332–346. [1575]

Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. (1999), "Missing Covariates in Generalized Linear Models When the Missing Data Mechanism Is Non-Ignorable," *Journal of the Royal Statistical Society*, Series B, 61, 173–190. [1575]

Im, K., Lee, J.-M., Seo, S. W., Kim, S. H., Kim, S. I., and Na, D. L. (2008a), "Sulcal Morphology Changes and Their Relationship With Cortical Thickness and Gyral White Matter Volume in Mild Cognitive Impairment and Alzheimer's Disease," *Neuroimage*, 43, 103–113. [1588,1592]

Im, K., Lee, J.-M., Seo, S. W., Yoon, U., Kim, S. T., Kim, Y.-H., Kim, S. I., and Na, D. L. (2008b), "Variations in Cortical Thickness With Dementia Severity in Alzheimer's Disease," *Neuroscience Letters*, 436, 227–231. [1588]

Jack, C. R., Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L Whitwell, J., Ward, C., and Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., and Weiner, M. W. (2008), "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods," *Journal of Magnetic Resonance Imaging*, 27, 685–691. [1586]

James, G. M. (2002), "Generalized Linear Models With Functional Predictors," *Journal of the Royal Statistical Society*, Series B, 64, 411–432. [1576]

Kang, J., Reich, B. J., and Staicu, A.-M. (2018), "Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process," *Biometrika*, 105, 165–184. [1583,1584,1591,1593]

Kim, J. K., and Yu, C. L. (2011), "A Semiparametric Estimation of Mean Functionals With Nonignorable Missing Data," *Journal of the American Statistical Association*, 106, 157–165. [1575]

Lee, S., Kawachi, I., Berkman, L. F., and Grodstein, F. (2003), "Education, Other Socioeconomic Indicators, and Cognitive Function," *American Journal of Epidemiology*, 157, 712–720. [1587,1588]

Li, T., Xie, F., Feng, X., Ibrahim, J. G., Zhu, H. T., and ADNI (2018), "Functional Linear Regression Models for Nonignorable Missing Scalar Responses," *Statistica Sinica*, 28, 1867–1886. [1576,1584]

Lièvre, A., Alley, D., and Crimmins, E. M. (2008), "Educational Differentials in Life Expectancy With Cognitive Impairment Among the Elderly in the United States," *Journal of Aging and Health*, 20, 456–477. [1587]

Lindley, D. V. (1972), *Bayesian Statistics: A Review* (Vol. 2), Philadelphia, PA: SIAM. [1577]

Ling, N., Liang, L., and Vieu, P. (2015), "Nonparametric Regression Estimation for Functional Stationary Ergodic Data With Missing at Random," *Journal of Statistical Planning and Inference*, 162, 75–87. [1576]

Little, R., and Rubin, D. (2002), *Statistical Analysis With Missing Data* (2nd ed.), Hoboken, NJ: Wiley. [1575]

Liu, T., Lipnicki, D. M., Zhu, W., Tao, D., Zhang, C., Cui, Y., Jin, J. S., Sachdev, P. S., and Wen, W. (2012), "Cortical Gyrification and Sulcal Spans in Early Stage Alzheimer's Disease," *PLoS One*, 7, e31083. [1588,1592]

McAuliffe, M. J., Lalonde, F. M., McGarry, D., Gandler, W., Csaky, K., and Trus, B. L. (2001), "Medical Image Processing, Analysis and Visualization in Clinical Research," in *14th IEEE Symposium on Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings*, IEEE, pp. 381–386. [1586]

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984), "Clinical Diagnosis of Alzheimer's Disease Report of the NINCDS-ADRDA Work Group* Under the Auspices of Department of Health and Human Services Task Force on Alzheimer's Disease," *Neurology*, 34, 939–939. [1588,1592]

Molenberghs, G., and Kenward, M. (2007), *Missing Data in Clinical Studies* (Vol. 61), New York: Wiley. [1575]

Morris, J. S. (2015), "Functional Regression," *Annual Review of Statistics and Its Application*, 2, 321–359. [1575]

Müller, H. G., and Stadtmüller, U. (2005), "Generalized Functional Linear Models," *The Annals of Statistics*, 33, 774–805. [1576]

Nestor, S. M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J. L., Fogarty, J., Bartha, R., and Alzheimer's Disease Neuroimaging Initiative (2008), "Ventricular Enlargement as a Possible Measure of Alzheimer's Disease Progression Validated Using the Alzheimer's Disease Neuroimaging Initiative Database," *Brain*, 131, 2443–2454. [1588,1592]

Persson, K., Bohbot, V., Bogdanovic, N., Selbæk, G., Brækhus, A., and Engedal, K. (2018), "Finding of Increased Caudate Nucleus in Patients With Alzheimer's Disease," *Acta Neurologica Scandinavica*, 137, 224–232. [1588]

Persson, K., Selbæk, G., Brækhus, A., Beyer, M., Barca, M., and Engedal, K. (2017), "Fully Automated Structural MRI of the Brain in Clinical Dementia Workup," *Acta Radiologica*, 58, 740–747. [1588]

Petersen, R. C., Roberts, R. O., Knopman, D. S., Geda, Y. E., Cha, R., Pankratz, V., Boeve, B. F., Tangalos, E. G., Ivnik, R. J., and Rocca, W. A. (2010), "Prevalence of Mild Cognitive Impairment Is Higher in Men the Mayo Clinic Study of Aging," *Neurology*, 75, 889–897. [1587]

Preda, C., Saporta, G., and Mbarek, M. H. B. H. (2010), "The NIPALS Algorithm for Missing Functional Data," *Revue Roumaine de Mathematique Pures et Appliquees*, 55, 315–326. [1576]

Qin, J., Leung, D., and Shao, J. (2002), "Estimation With Survey Data Under Nonignorable Nonresponse or Informative Sampling," *Journal of the American Statistical Association*, 97, 193–200. [1575]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, New York: Springer-Verlag. [1575]

Reiss, P. T., and Ogden, R. T. (2007), "Functional Principal Component Regression and Functional Partial Least Squares," *Journal of the American Statistical Association*, 102, 984–996. [1576]

——— (2010), "Functional Generalized Linear Models With Images as Predictors," *Biometrics*, 66, 61–69. [1576]

Shao, J., and Wang, L. (2016), "Semiparametric Inverse Propensity Weighting for Nonignorable Missing Data," *Biometrika*, 103, 175–187. [1575,1578,1588]

Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., and Leahy, R. M. (2001), "Magnetic Resonance Image Tissue Classification Using a Partial Volume Model," *NeuroImage*, 13, 856–876. [1586]

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998), "A Nonparametric Method for Automatic Correction of Intensity Nonuniformity in MRI Data," *IEEE Transactions on Medical Imaging*, 17, 87–97. [1586]

Smith, S. M. (2002), "Fast Robust Automated Brain Extraction," *Human Brain Mapping*, 17, 143–155. [1586]

Tang, G., Little, R. J., and Raghunathan, T. E. (2003), "Analysis of Multivariate Missing Data With Nonignorable Nonresponse," *Biometrika*, 90, 747–764. [1575]

Tang, N., Zhao, P., and Zhu, H. (2014), "Empirical Likelihood for Estimating Equations With Nonignorably Missing Data," *Statistica Sinica*, 24, 723–747. [1575]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [1579]

Visser, P. J., Verhey, F. R. J., Hofman, P. A. M., Scheltens, P., and Jolles, J. (2002), "Medial Temporal Lobe Atrophy Predicts Alzheimer's Disease in Patients With Minor Cognitive Impairment," *Journal of Neurology, Neurosurgery and Psychiatry*, 72, 491–497. [1588]

Wang, J. L., Chiou, J. M., and Mueller, H. G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295. [1575,1576]

Wang, S., Shao, J., and Kim, J. K. (2014), "An Instrumental Variable Approach for Identification and Estimation With Nonignorable Nonresponse," *Statistica Sinica*, 24, 1097–1116. [1575,1578]

Wang, X., and Zhu, H. (2017), "Generalized Scalar-on-Image Regression Models via Total Variation," *Journal of the American Statistical Association*, 112, 1156–1168. [1575,1583,1584,1593]

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Salazar, J., Saykin, A. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., and Alzheimer's Disease Neuroimaging Initiative (2017a), "The Alzheimer's Disease Neuroimaging Initiative 3: Continued Innovation for Clinical Trial Improvement," *Alzheimer's & Dementia*, 13, 561–571. [1574]

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., and Alzheimer's Disease Neuroimaging Initiative (2017b), "Recent Publications From the Alzheimer's Disease Neuroimaging Initiative: Reviewing Progress Toward Improved AD Clinical Trials," *Alzheimer's & Dementia*, 13, e1–e85. [1574]

Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., Petersen, R. C., and Jack, C. R. (2007), "3D Maps From Multiple MRI Illustrate Changing Atrophy Patterns as Subjects Progress From Mild Cognitive Impairment to Alzheimer's Disease," *Brain*, 130, 1777–1786. [1588]

Yang, F., Lorch, S. A., and Small, D. S. (2014), "Estimation of Causal Effects Using Instrumental Variables With Nonignorable Missing Covariates: Application to Effect of Type of Delivery NICU on Premature Infants," *The Annals of Applied Statistics*, 8, 48–73. [1575]

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [1576]

Yilmazer-Hanke, D. M., and Joachim, H. (1999), "Progression of Alzheimer-Related Neuritic Plaque Pathology in the Entorhinal Region, Perirhinal Cortex and Hippocampal Formation," *Dementia and Geriatric Cognitive Disorders*, 10, 70–76. [1588]

Yuan, M., and Cai, T. T. (2010), "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression," *The Annals of Statistics*, 38, 3412–3444. [1575]

Zhao, H., Zhao, P.-Y., and Tang, N.-S. (2013), "Empirical Likelihood Inference for Mean Functionals With Nonignorably Missing Response Data," *Computational Statistics & Data Analysis*, 66, 101–116. [1575]

Zhao, J., and Shao, J. (2015), "Semiparametric Pseudo Likelihoods in Generalized Linear Models With Nonignorable Missing Data," *Journal of the American Statistical Association*, 110, 1577–1590. [1575,1578]

Zhu, H., Fan, J., and Kong, L. (2014), "Spatially Varying Coefficient Model for Neuroimaging Data With Jump Discontinuities," *Journal of the American Statistical Association*, 109, 1084–1098. [1576]

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. (2011a), "Functional Principal Component Model for High-Dimensional Brain Imaging," *NeuroImage*, 58, 772–784. [1576,1580]

——— (2011b), "Multilevel Functional Principal Component Analysis for High-Dimensional Data," *Journal of Computational and Graphical Statistics*, 20, 852–873. [1580]