


Bayesian latent factor on image regression with nonignorable missing data

Xiaoqing Wang¹ | Xinyuan Song¹  | Hongtu Zhu²

¹Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong

²Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence

Xinyuan Song, Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong.
Email: xysong@sta.cuhk.edu.hk

Abstract

Medical imaging data have been widely used in modern health care, particularly in the prognosis, screening, diagnosis, and treatment of various diseases. In this study, we consider a latent factor-on-image (LoI) regression model that regresses a latent factor on ultrahigh dimensional imaging covariates. The latent factor is characterized by multiple manifest variables through a factor analysis model, while the manifest variables are subject to nonignorable missingness. We propose a two-stage approach for statistical inference. At the first stage, an efficient functional principal component analysis method is applied to reduce the dimension and extract useful features/eigenimages. At the second stage, a factor analysis mode is proposed to characterize the latent response variable. Moreover, an LoI model is used to detect influential risk factors, and an exponential tiling model applied to accommodate nonignorable nonresponses. A fully Bayesian method with an adjust spike-and-slab absolute shrinkage and selection operator (lasso) procedure is developed for the estimation and selection of influential features/eigenimages. Simulation studies show the proposed method exhibits satisfactory performance. The proposed methodology is applied to a study on the Alzheimer's Disease Neuroimaging Initiative data set.

KEYWORDS

imaging predictor, latent outcome, MCMC methods, nonignorable missingness, spike-and-slab lasso

1 | INTRODUCTION

Alzheimer's disease (AD), the most common cause of dementia, is an irreversible and progressive brain disorder that slowly destroys memory and thinking skills and, eventually, the ability to carry out the simplest tasks. It is the sixth leading cause of death in the United States. AD is characterized by the loss of neurons and synapses in the cerebral cortex and certain subcortical regions. A probable diagnosis can be acquired based on medical tests. However, initial AD symptoms are often mistaken for normal aging. A definite diagnosis requires brain tissue examination. Early diagnosis of AD is important. Although no cure has been found yet, treatments are available to slow the worsening of the symptoms and improve quality of life for patients with AD.

Neuroimaging is among the most promising areas of research focusing on early detection of AD because microscopic changes in the brain begin long before the first signs of memory loss. We focus on magnetic resonance imaging (MRI), a noninvasive structural imaging technique that can be acquired comparatively easily and has already been used extensively

for clinical diagnosis and medical research. Structural imaging provides information on the shape, position, or volume of brain tissue. Having shown that the brains of AD patients shrink significantly as the disease progresses, structural imaging research also has shown that shrinkage in specific brain regions, such as the hippocampus, may be an early sign of AD. However, scientists have not yet reached a consensus on this finding. In this study, we focus on MRI to explore its potential for use in the early diagnosis of AD.

Our study is motivated by the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set. The ADNI study began in 2004 and is the first "Big Data" project for AD. We aim to investigate the associations between the pathology of AD characterized by various cognitive tests and the MRI data as well as clinical/behavioral variables through a joint modeling approach, thereby providing new insights into the early diagnosis of AD. After applying a standard preprocessing pipeline, the dimension of the processed MRI images is $128 \times 128 \times 128$. The scores of three tests, namely, AD assessment scale 13 (ADAS13), Rey auditory verbal learning test-learning (RAVLT), and functional assessment questionnaire (FAQ), with over 40% missingness are available for the assessment of cognitive ability. Given the data background, the following scientific questions are of particular interest and worthy of investigation: (i) How can AD progression be measured using the three cognitive tests together? (ii) How can influential brain regions highly associated with AD progression be identified based on available MRI images? (iii) How can large amount of missingness of the cognitive test responses be managed effectively?

To answer the first question, we utilize a factor analysis model to formulate cognitive ability as a latent factor and measure it through the scores of the three tests. These test results are highly correlated and characterize cognitive ability from different perspectives. We consider formulating a latent response instead of implementing multiple regression because it can greatly reduce computational complexity and provide a clear model interpretation. Moreover, the latent variable retains the information that is useful for AD diagnosis from the multiple test scores. We use the doctor's diagnosis (CN or AD) as ground-truth to evaluate the K-means clustering results based on the multiple test scores and latent variable. The accuracy given by K-means is 76.8% based on the scores of ADAS13, RAVLT-learning, and FAQ, and 77.8% based on the latent variable (the value of the latent variable at the last MCMC iteration is used for simplicity). This result implies that the latent variable can adequately summarize the information of the multiple tests.

To answer the second question, we develop a latent factor-on-image (LoI) regression model to examine the association between cognitive decline and MRI images. Figure S1 in the Supplementary Material shows several slices of the processed MRI images from two randomly selected CN (the first two rows) and two randomly selected AD subjects (the last two rows). The MRI images exhibit distinct spatial features across CN and AD subjects. For example, compared to CN brains, the AD brains have enlarged ventricles (in the middle of the images, with dark color). Thus, incorporating image predictors with such strong spatial structures into a regression brings new challenges to statistical inference. These challenges include (i) the ultrahigh dimensional unknown parameter, which is of the same size as the three-dimensional (3D) image predictor, (ii) the complex spatial information, and (iii) the remarkable variability of brain structure across subjects. Many regularization methods, such as high-dimensional variable selection and supervised learning,^{1,2} are available for dimension reduction in the analysis of high-dimensional data. However, the unstructured regularization methods may suffer from diverging spectra and noise accumulation because of the strong spatial correlation in high-dimensional predictors and the nonsparse regression coefficients of these predictors.³ Meanwhile, the structured regularization methods (eg, the fused least absolute shrinkage and selection operator [lasso] or Ising prior) can be computationally challenging in the presence of ultrahigh-dimensional imaging predictors.⁴ Alternative methods for dimension reduction include principal component analysis (PCA) or screening methods, which extract a comparatively small number of important "features" or eliminate nonsignificant voxels.^{5,6} Among these methods, the functional PCA (FPCA) technique has shown high potential in feature extraction and prediction,^{7,8} where the image is regarded as functional data and the spatial association is wrapped up in the features/eigenimages. To our knowledge, the existing FPCA-related methods have focused on selecting the first several leading principal components with large variances.^{9,10} However, preceding studies argued that despite having small variances, later principal components could actually be more important than the leading ones in principal component regression, for example, References 11,12. Given these results, we propose a two-stage Bayesian approach to address the aforementioned challenges. In the first stage, we extract features/eigenimages from MRI images through (unsupervised) FPCA, which preliminarily reduces the dimensionality and spatial dependency of the imaging data. In the second stage, we introduce a spike-and-slab lasso (SSL) procedure¹³ to select significant features/eigenimages.

To address the third question, we must consider a plausible missing data mechanism. Figure S2 in the Supplementary Material displays the scatterplot of the baseline scores of the three tests for subjects with and without missingness at the 36th month follow-up visit. The missingness of the test scores do not appear to occur at random. Subjects with high scores of ADAS13 and FAQ or low scores of RAVLT are likely to be unresponsive in the follow-up visit. Such nonignorability of

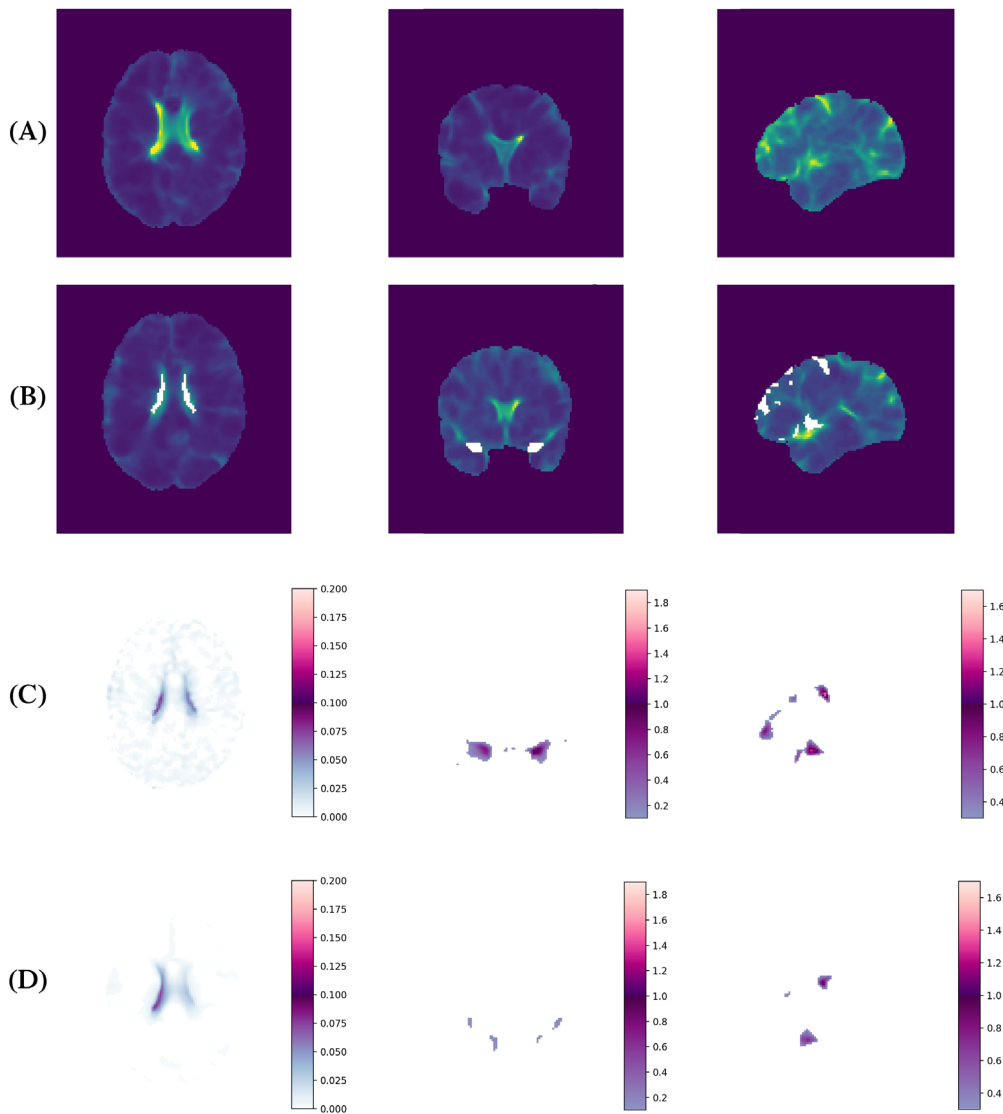


FIGURE 1 From left to right, S_1 , S_2 , S_3 : A, sample images; B, true image parameters β (white color); C, estimated $\hat{\beta}$ by adjusted SSL; D, estimated $\hat{\beta}$ by traditional BIC method [Colour figure can be viewed at wileyonlinelibrary.com]

missingness should be accommodated appropriately. Statistical inference in the presence of nonignorable missing data is a challenging problem because the missing mechanism is generally unknown and must be determined based on the data.^{14,15} In this study, we consider a logistic regression to model the nonignorable missingness of the three cognitive tests.

The present study contributes to the analysis of regression with imaging data in several aspects. First, we provide an efficient two-stage dimension reduction approach composed of a feature/eigenimage extraction through FPCA in the first stage and feature/eigenimage selection through SSL in the second stage. Unlike available methods that focus on several leading principal components with large variances, the proposed method also takes into account later principal components (PCs) with small variances. As far as we know, this is the first time to incorporate later PCs under FPCA framework. The simulation study presented in Section 4 (Figure 1C,D) demonstrates that the performance of the proposed method is significantly improved with the use of later PCs. The SSL procedure adopted in the second stage selects important features/eigenimages further by assigning spike-and-slab priors to the regression coefficients of LoI. Second, we introduce a latent factor to facilitate the integration of multiple measurements of cognitive ability through a factor analysis model. The introduction of the latent factor enables comprehensive characterization of cognitive ability and reduces the multivariate imaging regression to a univariate one, thereby facilitating a conceptually simple and attractive interpretation on the associations between cognitive ability and MRI images along with other potential risk factors. Finally, we propose the use of an exponential tilting model to accommodate the nonignorability of missingness in the context of the LoI regression model, thereby avoiding a biased inference caused by simple deletion/imputation of missing data.

The rest of the article is organized as follows. Section 2 presents the model description. Section 3 introduces the spike-and-slab prior and associated SSL procedures. In Section 4, simulation studies are conducted to examine the

performance of the proposed method. Section 5 applies the developed methodology to the ADNI study. Section 6 concludes the article. Technical details and additional results are provided in an online Supplementary Material.

2 | MODEL DESCRIPTION

2.1 | Functional principal component analysis (FPCA)

Suppose we have a sample of N 3D images of grid size (v_1, v_2, v_3) and $V = v_1 \times v_2 \times v_3$ is the total number of voxels. We vectorize the images before conducting FPCA. Let $\mathbf{X}_i = (X_i(1), \dots, X_i(V))'$, $i = 1, \dots, N$ denote the vectorized images. Following Zipunnikov et al⁷ and Di et al,¹⁶ we regard \mathbf{X}_i as functional data and consider a functional model $X_i(v) = \mu(v) + \tilde{X}_i(v)$, $i = 1, \dots, N$ where v denotes a voxel coordinate, image $\mu(v)$ is the overall mean image, and $\tilde{X}_i(v)$ is the subject-specific deviation from the overall mean. In what follows, we always work on $\tilde{X}_i(v)$. With a slight abuse of notation, we use $X_i(v)$ instead of $\tilde{X}_i(v)$ throughout this article to denote the voxelwise centered image. We regard $X_i(v)$ as a zero-mean second-order stochastic process with covariance function $\mathbf{K}(v_1, v_2) = \mathbb{E}[X_i(v_1)X_i(v_2)]$. As argued by Zipunnikov et al,⁷ the measurement errors of images are not included because the images are usually presmoothed (see a brief description in Section 5). The Karhunen-Loeve expansion of the centered random process¹⁷ has the form $X_i(v) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k(v)$, where $\phi_k(v)$ s are the eigenfunctions of the covariance function $\mathbf{K}(v_1, v_2)$, and ξ_{ik} are uncorrelated eigenscores with non-increasing variance σ_k^2 . Given N observations, we utilize a common estimator of the covariance operator \mathbf{K} , that is, $\hat{\mathbf{K}} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i \mathbf{X}_i'$. For imaging data, the number of voxels (V) is usually much larger than the sample size (N). Thus, we have $\text{rank}(\hat{\mathbf{K}}) = N$. The eigendecomposition of $\hat{\mathbf{K}}$ is $\hat{\mathbf{K}} = \hat{\mathbf{\Phi}}_N \hat{\mathbf{\Sigma}}_N \hat{\mathbf{\Phi}}_N'$ where $\hat{\mathbf{\Phi}}_N$ is a $V \times N$ matrix with orthonormal columns $\hat{\phi}_k$, which provide the estimates of eigenfunctions $\phi_k(v)$ s, and $\hat{\mathbf{\Sigma}}_N = \text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_N)$ is an $N \times N$ diagonal matrix with decreasing nonnegative eigenvalues.

Next, we consider an expansion of \mathbf{X}_i on N estimated eigenfunctions/eigenimages. With a slight abuse of terminology, we use eigenfunctions and eigenimages interchangeably throughout this article. The expansion is expressed as follows:

$$X_i(v) = \sum_{k=1}^N \hat{\xi}_{ik} \hat{\phi}_k(v) \quad \text{or} \quad \mathbf{X}_i = \hat{\mathbf{\Phi}}_N \hat{\xi}_i, \quad (1)$$

where $\hat{\xi}_i = (\hat{\xi}_{i1}, \dots, \hat{\xi}_{iN})'$, and $\hat{\xi}_{ik}$ is the k th eigenscore corresponding to the k th eigenimage $\hat{\phi}_k(v)$. Then, each image $\mathbf{X}_i = (X_i(1), \dots, X_i(V))'$ becomes a stacking of N eigenimages $\hat{\phi}_k = (\hat{\phi}_k(1), \dots, \hat{\phi}_k(V))'$, $k = 1, \dots, N$ with N corresponding uncorrelated eigenscores $\hat{\xi}_{i1}, \dots, \hat{\xi}_{iN}$. Thus, the dimension of the image is reduced from V to N . Meanwhile, the spatial dependencies of the elements of \mathbf{X}_i are described by the eigenimages.

Then, we show how to acquire the estimates of the eigenimages $\hat{\mathbf{\Phi}}_N$ and eigenscores $\hat{\xi}_i$, $i = 1, \dots, N$. In image data analysis, the number of voxels V can be extremely large, such as $128 \times 128 \times 128 = 2^{21}$. Conducting the eigendecomposition of a $V \times V$ matrix $\hat{\mathbf{K}}$ to acquire $\hat{\mathbf{\Phi}}_N$ is difficult. For example, the brute-force eigenanalysis requires $O(V^3)$ operation, which involves an infeasible level of complexity. Therefore, we consider an alternative approach to acquire $\hat{\mathbf{\Phi}}_N$. Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)$ be a $V \times N$ matrix. Notably, the number of images N is typically much smaller than the number of voxels V . Thus, the matrix \mathbf{X} has at most rank N , and the singular value decomposition (SVD) $\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}'$ can be obtained with $O(VN^2 + N^3)$ computational effort.¹⁸ \mathbf{U} is a $V \times N$ matrix with N orthonormal columns, \mathbf{S} is an $N \times N$ diagonal matrix, and \mathbf{V} is an $N \times N$ orthonormal matrix. Then, we show the relationship between SVD and FPCA. From the SVD, we have $\hat{\mathbf{K}} = \frac{1}{N} \mathbf{X} \mathbf{X}' = \frac{1}{N} \mathbf{U} \mathbf{S}^2 \mathbf{U}'$. Compare with the eigendecomposition of $\hat{\mathbf{K}}$, we have $\frac{1}{N} \mathbf{U} \mathbf{S}^2 \mathbf{U}' = \hat{\mathbf{\Phi}}_N \hat{\mathbf{\Sigma}}_N \hat{\mathbf{\Phi}}_N'$. Given that all eigenvalues of $\hat{\mathbf{K}}$ are different, the eigendecomposition of $\hat{\mathbf{K}}$ is unique. Thus, we acquire

$$\hat{\mathbf{\Phi}}_N = \mathbf{U}, \quad \hat{\mathbf{\Sigma}}_N = \frac{1}{N} \mathbf{S}^2. \quad (2)$$

Based on Equation (1), we have

$$\hat{\xi}_i = \mathbf{S} \mathbf{V}', \quad (3)$$

which is the estimated best linear unbiased predictors given by McCulloch et al.¹⁹

In summary, to obtain the estimated eigenimages $\hat{\Phi}_N$ and eigenscores $\hat{\xi}$ of FPCA, we only need to conduct SVD of the $V \times N$ matrix \mathbf{X} , and then apply Equations (2) and (3).

2.2 | Latent factor-on-image (LoI) regression model

First, we introduce a factor analysis model for characterizing a latent factor of interest. Consider a unidimensional latent factor η_i (eg, cognitive ability), which is characterized by a $q \times 1$ vector of observed variables $\mathbf{y}_i = (y_{i1}, \dots, y_{iq})'$ (eg, $q = 3$, and y_{i1}, y_{i2}, y_{i3} represent the three cognitive test scores). A factor analysis model is defined as follows:

$$\mathbf{y}_i = \boldsymbol{\mu} + \lambda \eta_i + \boldsymbol{\epsilon}_i, \tag{4}$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_q)'$ is a $q \times 1$ vector of intercepts, $\lambda = (\lambda_1, \dots, \lambda_q)'$ is a $q \times 1$ vector of factor loadings, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{iq})'$ is a $q \times 1$ vector of random errors independent of η_i and distributed as $N(\mathbf{0}, \boldsymbol{\Psi}_\epsilon)$ with a diagonal covariance matrix $\boldsymbol{\Psi}_\epsilon$. Notably, the factor analysis model (4) is unidentifiable because multiplying λ simultaneously by an arbitrary nonzero constant C and η_i by $\frac{1}{C}$ results in the same likelihood function, implying that λ and η_i are not simultaneously estimable. This indeterminacy can be addressed by fixing appropriate elements in λ at preassigned values, for example, References 20,21. In the present study, we fix the first element in λ at 1.

Let \mathbf{z}_i be a $r_1 \times 1$ vector of observed covariates and \mathbf{X}_i be the centered and vectorized image predictor. We define an LoI regression model as follows:

$$\eta_i = \boldsymbol{\alpha}' \mathbf{z}_i + \frac{1}{c_V} \sum_{v=1}^V \beta(v) X_i(v) + \delta_i, \tag{5}$$

where η_i is the latent factor defined in (4), $\boldsymbol{\alpha}$ is a $r_1 \times 1$ vector of regression coefficients; $\beta(v)$ is the regression coefficient of $X_i(v)$ at the v th voxel; c_V is a normalizing scalar determined by the total number of voxels V ; δ_i is a random error; and $\delta_i \sim N(0, \psi_\delta)$. The interpretation of $\boldsymbol{\beta} = (\beta(1), \dots, \beta(V))'$ is that the image regions with large $|\beta(\cdot)|$ are influential. The normalizing scalar c_V can be absorbed into the image predictors. The role of c_V is to rescale the total effect of massive image predictors (ie, the summation in (5)), such that this total effect is bounded away from infinity when V increases. The value of c_V can be chosen as $V^{1/2}$.²²

For the ultrahigh-dimensional centered image vector $\mathbf{X}_i = (X_i(1), \dots, X_i(V))'$, the FPCA is applied as described in the previous session to reduce dimensionality. With the eigenimages $\hat{\Phi}_N = (\hat{\phi}_1, \dots, \hat{\phi}_N)$ derived and estimated from the FPCA, \mathbf{X}_i has unique Karhunen-Loeve expansion $X_i(v) = \sum_{k=1}^N \hat{\xi}_{ik} \hat{\phi}_k(v)$. We assume that $\boldsymbol{\beta} = (\beta(1), \dots, \beta(V))'$ can also be expanded on $\hat{\Phi}_N$. While this assumption restricted the flexibility of $\beta(\cdot)$, it can simplify the problem significantly. We demonstrate in Section 5 that a satisfactory result can be obtained even such an assumption is violated. Under this assumption, we have

$$\beta(v) = \sum_{k=1}^N \beta_k \hat{\phi}_k(v), \tag{6}$$

where β_k is the coefficient of the k th eigenimage $\hat{\phi}_k(v)$. The orthonormality of the eigenimages implies that $\sum_{v=1}^V \hat{\phi}_k^2(v) = 1$ and $\sum_{v=1}^V \hat{\phi}_k(v) \hat{\phi}_l(v) = 0$, for $k \neq l$. Thus, we have $\sum_{v=1}^V \beta(v) X_i(v) = \sum_{k=1}^N \beta_k \hat{\xi}_{ik}$.

In substantive study, the last portion of eigenimages can be very massy and of little information. Thus, we follow the common practice of incorporating the first P eigenimages that keep approximately 99% of the variability $[(\hat{\sigma}_1 + \dots + \hat{\sigma}_P) / (\hat{\sigma}_1 + \dots + \hat{\sigma}_N) = 0.99]$.²³ The LoI regression in Equation (5) can be rewritten as follows:

$$\eta_i = \boldsymbol{\alpha}' \mathbf{z}_i + \frac{1}{c_V} \sum_{k=1}^P \beta_k \hat{\xi}_{ik} + \delta_i, \tag{7}$$

where $\hat{\xi}_{ik}$ is the k th eigenscore estimated in the first stage and β_k is the unknown coefficient of $\beta(v)$ expansion. Once we acquire the estimation of β_k , we can use Equation (6) (with the first P eigenimages) to reconstruct $\beta(v)$. Although P is much smaller than V , it is still comparable to sample size N because the decreasing of variance is slow in FPCA.²⁴

2.3 | Missing data mechanism

We introduce an indicator vector $\mathbf{r}_i = (r_{i1}, \dots, r_{iq})$ to label the missing entries in $\mathbf{y}_i = (y_{i1}, \dots, y_{iq})'$, where $r_{ij} = 1$ if y_{ij} is missing and $r_{ij} = 0$ otherwise. We assume a Bernoulli distribution of r_{ij} : for $j = 1, \dots, q$, $r_{ij}|y_{ij}, \mathbf{w}_i \sim \text{Bernoulli}(\pi_j)$, where \mathbf{w}_i is an $r_2 \times 1$ vector of covariates. Let $p_{ij} = \text{Pr}(r_{ij} = 1|y_{ij}, \mathbf{w}_i)$. A missing data model is considered as follows:

$$\text{logit}(p_{ij}) = \rho_{j0} + \rho'_{j1} \mathbf{w}_i + \varphi_j y_{ij}, \tag{8}$$

where ρ_{j0} is an intercept, ρ_{j1} is an $r_2 \times 1$ vector of regression coefficients, and φ_j is the tilting parameter that determines the amount of departure from the ignorability of the missing mechanism. If $\varphi_j = 0$, then the missing probability does not depend on the missing data, and the missing mechanism becomes ignorable.

3 | POSTERIOR INFERENCE

3.1 | Spike-and-slab lasso (SSL) prior

In Equation (7), the number of unknown parameters is $(r_1 + P)$, which is comparable to sample size N . We assume the sparsity of $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)'$ because of the following reasons. First, the eigenimages can be considered as features extracted from the imaging data. Naturally, not all features are important for explaining the variation of the images. Second, this assumption is common for a feasible statistical inference in high-dimensional regression and variable selection problems.

To select and estimate significant β_k s under a high-dimensional setting, we consider the spike-and-slab lasso prior¹³ as follows:

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{k=1}^P [\gamma_k \psi_1(\beta_k) + (1 - \gamma_k) \psi_0(\beta_k)], \tag{9}$$

where $\psi_1(\beta) = \frac{\nu_1}{2} e^{-\nu_1 |\beta|}$ with small ν_1 serves as a diffused “slab distribution” for modeling large effects; $\psi_0(\beta) = \frac{\nu_0}{2} e^{-\nu_0 |\beta|}$ with large ν_0 serves as a concentrated “spike distribution” for modeling negligibly small effects; and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_P)'$, $\gamma_k \in \{0, 1\}$ indexes the 2^P possible subset models. The scope of SSL priors can be vastly enhanced with different choices of $p(\boldsymbol{\gamma})$. Inspired by Ročková and George,¹³ we consider an exchangeable prior for $p(\boldsymbol{\gamma})$ as follows:

$$p(\boldsymbol{\gamma}|\pi) = \prod_{k=1}^P \pi^{\gamma_k} (1 - \pi)^{1-\gamma_k}, \quad p(\pi) = \text{Beta}(a_1, b_1), \tag{10}$$

where $\pi = P(\gamma_k = 1)$ is the expected fraction of large β_k s. Through this π parameter, an opportunity to learn about the level of sparsity of $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)'$ is provided to the penalty imposed on the regression coefficients. A conjugate Beta prior is assigned to π . we choose hyperparameter $a_1 = 1$ and $b_1 = CP$ for some constant $C > 0$ to ensure the posterior consistency of the parameter estimators.¹³

We also consider a variant of SSL. The abovementioned SSL priors assume that all predictors are centered and standardized. However, the variances of eigenscores ξ_{ik} in Equation (7) decrease as k increase, leading to a violation of the standardization assumption. Thus, we consider an adjusted “spike distribution” as follows:

$$\psi_{0k}(\beta_k) = \frac{\nu_0 \tau_k}{2} \exp(-\nu_0 \tau_k |\beta_k|), \quad k = 1, \dots, P. \tag{11}$$

The ψ_{0k} is now a Laplace distribution with parameter $\nu_0 \tau_k$, where τ_k is a parameter for adjusting ν_0 across k to cope with the decreasing variance of ξ_{ik} , $k = 1, \dots, P$. We impose an inverse-Gamma prior for τ_k as follows:

$$p(\tau_k) = \text{IGamma}(a_{2k}, b_{2k}), \quad k = 1, \dots, P, \tag{12}$$

where hyperparameters a_{2k} and b_{2k} are chosen as $a_{2k} = \frac{1}{\text{sd}(\xi_{ik})} + 1$ and $b_{2k} = 1$, and $\text{sd}(\xi_{ik})$ is the standard deviation of ξ_{ik} , $i = 1, \dots, N$. Consequently, the prior distribution of τ_k has a mean $\text{sd}(\xi_{ik})$ and small variance. Given that the posterior

samples of $\tau_k v_0$ tend to be small if $\text{sd}(\xi_{ik})$ is small, β_k is more likely to be shrunk to zero.¹³ The performance of the SSL prior and the adjusted version (Equation (11)) will be evaluated through simulation studies in Section 4.

For other parameters, we assign commonly used prior distributions, most of which are conjugate-type priors. The details are provided in the Supplementary Material.

3.2 | Posterior sampling

Let \mathbf{D} be the set of observed data and θ be the vector of all unknown parameters including $\mu, \lambda, \Psi_e, \alpha, \beta, \gamma, \pi, \psi_\delta, \rho,$ and ϕ . The main task of Bayesian inference is to estimate θ by sampling from its posterior distribution $p(\theta|\mathbf{D})$. However, this posterior distribution is intractable because it involves high-dimensional integrations due to the existence of latent variable η_i and missing data in \mathbf{y}_i . Thus, we employ the idea of data augmentation. Let $\eta = \{\eta_i : i = 1, \dots, N\}$ be the set of latent variables, $\mathbf{y}_m = \{y_{ij} | r_{ij} = 1, i = 1, \dots, N; j = 1, \dots, q\}$ be the set of missing data. Then, \mathbf{D} is augmented with η and \mathbf{y}_m , and the Bayesian estimate of θ can be obtained by using the samples drawn from the joint posterior distribution $p(\theta, \eta, \mathbf{y}_m | \mathbf{D})$. We use Gibbs sampler to sample iteratively from the full conditional distributions $p(\mathbf{y}_m | \theta, \eta, \mathbf{D}), p(\eta | \theta, \mathbf{y}_m, \mathbf{D}),$ and $p(\theta | \mathbf{y}_m, \eta, \mathbf{D})$. Some of the abovementioned posterior distributions are not in familiar forms because of the nonignorable missing mechanism and the SSL prior. Hence, the Metropolis-Hastings algorithm is utilized to sample from these complex distributions. We determine convergence of the algorithm by examining the plots of several MCMC chains starting from different initial values. At convergence, these MCMC chains should mix adequately. The full conditional distributions involved in the Gibbs sampler are provided in the Supplementary Material.

4 | SIMULATION STUDY

In this simulation, we examine the performance of the proposed method for an LoI model with 2D images (128×128). The images are 2D slices extracted from the preprocessed $N = 806$ 3D ($128 \times 128 \times 128$) MRI images in the ADNI data set. The details of the preprocessing procedures will be explained in the real data analysis section. We consider three scenarios of 2D image predictors (\mathbf{X}_i) as follows:

- Scenario 1 (S1): axial slice with thalamus,
- Scenario 2 (S2): coronal slice with hippocampus, and
- Scenario 3 (S3): sagittal slice with frontal lobe gyrus.

Three sample images generated under the above scenarios are shown in the first row of Figure 1A. For each scenario, a true parameter image $\beta = (\beta(1), \dots, \beta(V))'$ is specified. The shapes of the true parameters are shown in Figure 1B (in white color), which can be interpreted as the thalamus, hippocampus, and sulcus in the frontal lobe, respectively. The true values at voxels inside the specified shape are all set to be 0.1 for S1, 1.0 for S2 and S3, and the true values at other voxels are all set to be 0. The other parts of the model are the same across all the scenarios. For the LoI model (Equation (7)), let $r_1 = 2, \mathbf{z}_i = (z_{i1}, z_{i2})'$, where z_{i1} and z_{i2} are generated independently from $N(0, 1), \alpha = (0.8, 0.8)'$, and $\psi_\delta = 0.3$. For the factor analysis model (Equation (4)), let $q = 4, \mu = (0, 0, 0, 0)', \lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)' = (1.0^*, 0.8, 0.8, 0.8)',$ where the first element is fixed at 1.0 for model identification, and $\psi_{\epsilon 1} = \psi_{\epsilon 2} = \psi_{\epsilon 3} = \psi_{\epsilon 4} = 0.3$. For the exponential tilting model (Equation (8)), let $r_2 = 2, \mathbf{w}_i = (w_{i1}, w_{i2})'$, where w_{i1} and w_{i2} are generated independently from $N(0, 1), \rho_{j0} = -2, \rho_{j1} = (0.5, 0.5)',$ and $\psi_j = 0.5$. The overall missing proportion of \mathbf{y}_i is approximately 40%.

We use scenario 1 as an example to demonstrate the procedure. After extracting 2D images $\mathbf{X}_i, i = 1, \dots, N$ from the MRI images of the ADNI data set, we generate η_i according to Equation (5), y_{ij} according to Equation (4), and r_{ij} according to Equation (8). Then, we conduct the two-stage analysis. In the first stage, we estimate the eigenimages and eigenscores using FPCA. Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)$ be the $V \times N$ imaging data matrix. Here, $V = 4854$ (instead of 128^2) because we delete the background voxels with values always being 0. We can obtain the SVD of \mathbf{X}/c_V (the constant $c_V = \sqrt{V} \approx 70$ is absorbed into the image covariates) using Matlab function “svd.” The eigenimages $\hat{\Phi}_N$ and eigenscores $\hat{\xi}$ are obtained based on Equations (2) and (3). Figure S3(a) of the Supplementary Material shows the first eight estimated eigenimages.

In the second stage, we conduct Bayesian inference using SSL and adjusted SSL. Notably, instead of incorporating all N eigenimages, we actually only use the first $P = 380$ eigenimages that explain 99.05% of the variability defined

by $(\hat{\sigma}_1 + \dots + \hat{\sigma}_P)/(\hat{\sigma}_1 + \dots + \hat{\sigma}_N)$, where $\hat{\sigma}_k$ s are the eigenvalues that can be acquired from $\hat{\Sigma}_N = \frac{1}{N}\mathbf{S}^2$. One reason for using the first $P=380$ eigenimages is that the variances of later eigenscores are all less than 0.002, which contains very little information. Another reason is that the eigenimages corresponding to later eigenscores become very messy (the 381th–388th eigenimages are shown in Figure S3(b) of the Supplementary Material). Given that our purpose is to reconstruct β by selecting the desired features among these eigenimages, leaving out the last portion of eigenimages is reasonable.

In the implementation of the SSL and adjusted SSL procedures, we initially specify the hyperparameters. For the beta prior of π (Equation (10)), we use $a_1 = 1$ and $b_1 = P$ to introduce the sparsity to the regression coefficients as suggested by Ročková and George.¹³ For the prior of τ_k , $k = 1, \dots, P$ in the adjusted spike distribution (Equation (12)), we use $a_{2k} = \left(\frac{1}{\text{sd}(\xi_k)} + 1\right)_{[3,10]}$ and $b_{2k} = 1$, and let $3 \leq a_{2k} \leq 10$, so that the mean $\frac{b_{2k}}{a_{2k}-1}$ of the inverse-Gamma prior will not become too small and the variance $\frac{b_{2k}}{(a_{2k}-1)^2(a_{2k}-2)}$ will not explode. The rest of hyperparameters are chosen to formulate vague prior information. For the spike-and-slab penalty, we follow the suggestion of Ročková and George¹³ to assign $v_1 = 1$ and $v_0 = P^d$, where d is a positive constant. Notably, $v_0 = P^d$ only gives a reference of choosing v_0 . The specific value of v_0 is determined by using the modified BIC²⁵ as follows: $\text{BIC}_S = \log(\hat{\sigma}_S^2) + |S| \frac{\log N}{N} C_N$, where N is the sample size, $|S|$ denotes the size of the model, $\hat{\sigma}_S^2 = \text{SSE}_S/N$, and $C_N \rightarrow \infty$ with an arbitrary slow rate. Following the suggestion of Wang et al,²⁵ we take $C_N = \log \log P$, where P is the number of parameters comparable to the sample size N .

Estimated potential scale reduction (EPSR) proposed by Gelman et al²⁶ is used to check the convergence of algorithm. Some pilot runs show that the MCMC algorithm converges within 8000 iterations. Thus, we collect 10 000 simulated observations after 8000 burn-ins to obtain the Bayesian estimates of the unknown parameters. The computation time for a single replication is 0.41 hours on Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80 GHz. β_k is considered to be different from zero if the proportion of the posterior samples of $\gamma_k = 1$ is over 50%. The image parameter $\beta = (\beta(1), \dots, \beta(V))'$ is then reconstructed according to Equation (6) by using the estimated β_k , $k = 1, \dots, P$. The performance of the proposed method is evaluated based on 100 replications under all scenarios. For the adjusted SSL, the number of selected eigenimages P_0 , and the bias (Bias) and root mean square error (RMSE) between the Bayesian estimates and true population values of parameters are presented in Table 1. For the image parameter β , we present the average Bias and RMS across all voxels. The reconstructed estimated image parameters β are shown in Figure 1C. The adjusted SSL procedure performs satisfactorily in terms of bias and RMS for all scenarios, and both the shapes and values of the true parameter images β are well captured. On the contrary, based on the modified BIC, the nonadjusted SSL procedure selects a model without any eigenimage. Thus, the result corresponding to the SSL is unsatisfactory and not presented here. We also consider implementing the traditional method, which uses BIC to select several leading eigenimages. Under S1, S2, and S3, the first 5, 3, and 8 eigenimages are selected, respectively. The reconstructed parameter images β are shown in Figure 1D, which are unfavorable.

Notably, in the above simulation study, although the true parameter images $\beta(v)$ are chosen based completely on practical meaning rather than by verifying the assumption on whether they can be expanded on $\hat{\Phi}$, the proposed method can still give a good recovery of the true value and shape.

The performances of SSL and adjusted SSL are further compared through simulation 2 focusing on the second-stage model. The results show that the adjusted SSL performs much better than SSL. We also investigated the difference of 90%, 95%, and 99% criteria in choosing p and find that the estimation results are stable. The details of simulation 2 are provided in the Supplementary Material.

5 | A STUDY ON THE ADNI DATA SET

We apply the proposed method to the ADNI data set described in the Introduction. Our main interest is to examine the association between the cognitive ability measured by the three clinical cognitive test scores at the 36th month follow-up visit and the MRI images along with several covariates obtained during the baseline screening visit. The three cognitive tests that we focused on are ADAS13 (y_1), RAVLT (y_2), and FAQ (y_3). Their scores are highly correlated and subject to 45.78% missingness. The observed test scores were standardized prior to analysis. The RAVLT scores were negated, such that large scores correspond to bad cognitive conditions. The MRI data, which were collected across a variety of 1.5 Tesla MRI scanners with protocols individualized for each scanner, included standard T1-weighted images obtained using volumetric 3D sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol included the following variables: repetition time (TR) = 2400 milliseconds, inversion time (TI) = 1000 milliseconds, flip angle = 8°, and

TABLE 1 Bayesian estimation of parameters in simulation 1 with three scenarios, S1, S2, and S3

Number of selected eigenimages							
	S1	S2	S3				
P_0	6.523	57.418	58.523				
Bias (RMSE)				Bias (RMSE)			
Para	S1	S2	S3	Para	S1	S2	S3
μ_1	-0.015 (0.020)	0.007 (0.018)	0.012 (0.026)	λ_1	—	—	—
μ_2	-0.014 (0.021)	0.004 (0.016)	0.010 (0.019)	λ_2	-0.002 (0.006)	-0.001 (0.003)	-0.000 (0.002)
μ_3	-0.016 (0.021)	0.003 (0.016)	0.004 (0.018)	λ_3	-0.002 (0.006)	-0.001 (0.003)	-0.001 (0.003)
μ_4	-0.013 (0.018)	0.003 (0.015)	0.007 (0.020)	λ_4	-0.002 (0.006)	-0.002 (0.003)	-0.001 (0.002)
α_1	-0.009 (0.023)	0.002 (0.025)	-0.000 (0.034)	α_2	-0.007 (0.021)	-0.003(0.027)	0.006 (0.041)
ψ_{e1}	-0.164 (0.136)	-0.159 (0.159)	-0.154 (0.155)	ψ_{e3}	-0.171 (0.172)	-0.170 (0.170)	-0.165 (0.165)
ψ_{e2}	-0.173 (0.173)	-0.171 (0.171)	-0.166 (0.167)	ψ_{e4}	-0.172 (0.173)	-0.169 (0.169)	-0.166 (0.166)
ψ_δ	0.020 (0.037)	0.186 (0.140)	0.414 (0.335)	π	—	—	—
ρ_{10}	-0.045 (0.165)	0.006 (0.139)	-0.054 (0.206)	ρ_{11}	0.022(0.104)	-0.015 (0.117)	0.025 (0.123)
ρ_{20}	-0.024 (0.128)	-0.068 (0.145)	-0.070(0.172)	ρ_{21}	0.023(0.115)	0.012 (0.101)	0.024 (0.125)
ρ_{30}	-0.035 (0.134)	-0.072 (0.168)	-0.071(0.184)	ρ_{31}	-0.001(0.097)	0.013 (0.118)	0.024 (0.145)
ρ_{40}	-0.007 (0.137)	-0.079(0.180)	-0.055(0.178)	ρ_{41}	0.020(0.113)	0.000 (0.122)	0.015 (0.109)
ρ_{12}	-0.001 (0.111)	-0.001(0.119)	0.000 (0.130)	φ_1	0.022 (0.051)	0.006 (0.036)	0.016 (0.044)
ρ_{22}	0.029 (0.112)	0.006 (0.112)	0.031 (0.135)	φ_2	0.006 (0.055)	0.017 (0.043)	0.019 (0.040)
ρ_{32}	0.018 (0.109)	0.029 (0.117)	0.014 (0.122)	φ_3	0.009 (0.060)	0.018 (0.045)	0.011 (0.045)
ρ_{42}	0.002 (0.100)	0.019 (0.130)	0.015 (0.165)	φ_4	0.000 (0.053)	0.024 (0.058)	0.006 (0.040)

Note: P_0 is the average number of selected eigenimages by adjusted SSL based on 100 replications.

field of view (FOV) = 24 cm with a $256 \times 256 \times 170$ acquisition matrix in the x -, y -, and z -dimensions yielding a voxel size of $1.25 \times 1.26 \times 1.2 \text{ mm}^3$.²⁷ The MRI data were preprocessed by standard steps including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, registration,²⁸⁻³¹ and finally down-sampled to $128 \times 128 \times 128$ for our real data analysis. The other covariates under consideration include one genetic variable, namely, the APOE4 gene (z_{i1}), which is the only known genetic determinant of AD; $z_{i1} = 1$ if the subject carries at least one APOE4 allele and $z_{i1} = 0$ otherwise. In addition, gender (z_{i2} , 1 = male, 0 = female), age (z_{i3}), education level (z_{i4} : total years of education), race (z_{i5} , 1 = white, 0 = other), and marriage status (z_{i6} , never married = 1, ever married = 0) were included in the LoI model. In the exponential tilting model, we chose the covariates $\mathbf{w}_i = (z_{i1}, \dots, z_{i5})'$ to examine their effects on the probability of missingness of \mathbf{y}_i .

A total of 832 subjects with baseline 1.5T MRI scans acquired from <http://adni.loni.usc.edu/> were collected and pre-processed. The three test scores and six covariates were extracted from the “key ADNI tables merged into one table” from the same website. By matching the subject ID, we obtained 807 subjects who have both baseline MRI images and clinical records. After deleting a subject with unknown marriage status, we had $N = 806$ subjects (225 CN, 394 MCI, 187 AD) with an average age 75.24 of whom 470 are males and 336 are females. For each subject, the observed data included $128 \times 128 \times 128$ MRI images (\mathbf{X}_i), three test scores (\mathbf{y}_i), missing indicator vector (\mathbf{r}_i), and covariates (\mathbf{z}_i , \mathbf{w}_i).

For our analysis, we applied the proposed two-stage approach. In the first stage, we deleted the background voxels of the $128 \times 128 \times 128$ preprocessed MRI images to reduce the total number of voxels from 2^{21} to $V = 201\,337$. Then, the remaining voxels are voxelwise centered and $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)$ is a $201\,337 \times 806$ data matrix composed of vectorized centered nonbackground voxels. We conduct SVD of \mathbf{X}/c_V ($c_V = \sqrt{V} \approx 449$), which can be accomplished easily and quickly using the Matlab function “svd.” The eigenimages and eigenscores are then obtained using Equations (2) and (3). Notably, the spatial structure and correlation of the brain is reasonably addressed by the eigenimages. For example, the first eigenimage mainly describes the central ventricle, and the second eigenimage mainly accounts for the cortex (slices

TABLE 2 Bayesian estimation of parameters in the analysis of ADNI data set with $\nu_0 = 14$

Parameters in the LoI regression model						Other parameters					
Para	Est	SD	Para	Est	SD	Para	Est	SD	Para	Est	SD
α_1	0.533	0.128	β_1	-0.222	0.040	μ_1	2.000	0.372	λ_1	—	—
α_2	0.252	0.177	β_2	-0.478	0.060	μ_2	1.700	0.317	λ_2	0.831	0.039
α_3	-0.519	0.085	β_3	-0.865	0.106	μ_3	1.869	0.355	λ_3	0.949	0.035
α_4	-0.277	0.068	β_4	-0.642	0.120	φ_1	3.866	0.555	ψ_{ϵ_1}	0.264	0.028
α_5	0.060	0.276	β_5	-1.187	0.148	φ_2	5.618	0.546	ψ_{ϵ_2}	0.597	0.045
α_6	-0.803	0.256	β_6	-0.527	0.169	φ_3	2.976	0.325	ψ_{ϵ_3}	0.366	0.033
			β_7	0.606	0.174	ρ_{10}	-7.218	1.624	ρ_{11}	0.281	0.265
ψ_δ	2.496	0.265	β_8	-0.426	0.154	ρ_{20}	-7.775	1.276	ρ_{21}	0.230	0.269
π	0.016	0.004	β_{11}	1.022	0.185	ρ_{30}	-4.776	0.990	ρ_{31}	0.203	0.217
			β_{16}	0.678	0.197	ρ_{12}	0.823	0.596	ρ_{13}	-0.230	0.251
			β_{17}	0.636	0.201	ρ_{22}	0.082	0.565	ρ_{23}	-0.073	0.284
			β_{35}	0.922	0.257	ρ_{32}	-0.279	0.441	ρ_{33}	-0.092	0.216
						ρ_{14}	0.502	0.798	ρ_{15}	-1.198	0.506
						ρ_{24}	-0.804	1.110	ρ_{25}	-0.507	0.543
						ρ_{34}	-0.107	0.940	ρ_{35}	-1.084	0.399

of the first and second eigenimages are shown in Figure S6(a,b) of the Supplementary Material). In the second stage, instead of using all $N = 806$ eigenscores, we only incorporated the first $P = 430$ eigenscores that explain approximately 90% of the total variability³² because the later eigenimages became massy. Slices of the 431th and 432th eigenimages are shown in Figure S6(c,d) of the Supplementary Material. We applied the SSL and adjusted SSL to conduct the subsequent analysis. The priors were specified similar to that of Simulation 1. We first assigned $\nu = 1$ and $\nu_0 = P^d$ and then determined the specific value of ν_0 by using the modified BIC. The EPSR plot depicted in Supplementary Material Figure S8 showed the MCMC algorithm converged within 10 000 iterations. Therefore, we discarded the first 10 000 iterations as burn-ins and collected the subsequent 10 000 iterations for Bayesian inference. The computation time was 11.62 hours on Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80 GHz. For the SSL prior, regardless of the choice of ν_0 , obtaining a reliable result is difficult. Actually, when ν_0 was set to a small value, say, 3.0, the SSL procedure selected a model without any eigenimage. When ν_0 was slightly increased, however, the result jumped to a model of considerably large size with the variance (ψ_δ) exploded. On the contrary, the adjusted SSL procedure produced reasonable results. Thus, we only present and discuss the result of the adjusted SSL in the following interpretations. Table 2 reports the parameter estimates, and selected slices of the reconstructed estimated image parameter are shown in Figure S7 of the Supplementary Material. Based on the obtained results, we have the following observations.

In the factor analysis model, all factor loadings differ substantially from zero ($\hat{\lambda} = (1^*, 0.831, 0.949)'$), thereby indicating that the latent factor “cognitive ability (η_i)” is well characterized by the three tests. In the LoI regression model, the image parameter $\hat{\beta}$ was reconstructed. We detected several key regions of interest, which are in good agreement with the existing literature. For example, the positive effects of the hippocampal and amygdalar regions on cognitive ability are identified (Figure 2A). Among the core biomarkers of AD, hippocampal atrophy is the most validated and has been used in research to stage the progression of AD pathology in the brain across the entire disease spectrum.³³ The identified positive effect of the lateral temporal and dorsal parietal cortices (Figure 2B) is consistent with the literature that neocortical atrophy is related closely to AD progression.³⁴ A reduction in the volumes of the putamen and thalamus is found to have a strong positive effect on cognitive decline (Figure 2C), which agrees with the literature that in addition to neo-cortical atrophy, deep gray matter structures, such as hippocampus, putamen, and thalamus, can also contribute to AD progression.³⁵ Negative effects of medial frontal gyrus as well as perirhinal and entorhinal cortices are detected (Figure 2D), which align with the findings in neuroscience studies that AD severity is negatively associated with atrophy or cortical thickness in the frontal and hippocampal regions.³⁶

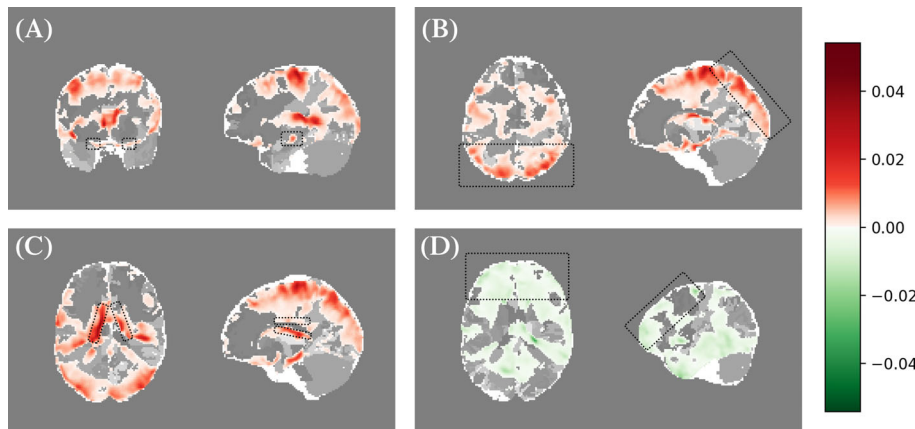


FIGURE 2 A, hippocampal and amygdalar; B, lateral temporal and dorsal parietal cortices; C, putamen and thalamus; D, frontal gyrus, perirhinal, and entorhinal cortices [Colour figure can be viewed at wileyonlinelibrary.com]

In addition to being an image predictor, APOE4 contributes to a worse cognitive condition ($\hat{\alpha}_1 = 0.533(0.128)$), whereas higher education level ($\alpha_4 = -0.277(0.068)$) and married ($\alpha_6 = -0.803(0.265)$) are associated with better cognitive condition. These findings are in line with existing medical literature.³⁷ However, age was also detected to contribute to better cognitive condition. A possible reason could be that the ADNI study focused on the elderly population (around 70 years old), and healthy elder people tend to possess better cognitive ability. Further investigation is required to comprehend better this unexpected result. In the exponential tilting model, race had negative effects on the missing probability of ADAD13 ($\hat{\rho}_{15} = -1.198(0.506)$) and FAQ ($\hat{\rho}_{35} = -1.084(0.399)$). The tilting parameters are all positive and differ substantially from zero ($\hat{\phi}_1 = 3.866(0.555)$, $\hat{\phi}_2 = 5.618(0.546)$, $\hat{\phi}_3 = 2.976(0.325)$), thereby reconfirming the nonignorability of the missingness.

6 | DISCUSSION

We proposed an LoI regression model for examining the effects of imaging and other predictors on a latent construct of interest in the presence of nonignorable missing data. A two-stage method composed of a FPCA method and SSL-related procedures was developed to perform statistical inference. Simulation studies showed that our method can recover the true parameters satisfactorily, including those related to the imaging part. The real data analysis results identified important brain regions associated with AD pathology.

In the present study, the outcome is cognitive ability, which is measured by three cognitive tests. The traditional method establishes a regression model for each test outcome and then relates the three test outcomes by assuming a covariance matrix for their measurement errors. By contrast, the proposed approach introduced a factor analysis model to group the three test outcomes into a latent factor and then regressed the latent factor on imaging predictors and other covariates. Consequently, the proposed approach reduces the multivariate scalar-on-image regression problem to a univariate LoI regression. The model dimensionality and computational burden are alleviated considerably. Meanwhile, the interpretations of model parameters become simple and comprehensible. Such a joint modeling approach also has high potential to be generalized to neuroimaging/genomics studies.³⁸

The present study has several limitations. First, our method was implemented in a two-stage basis. A single-stage method, such as Ising and Gaussian process priors,^{22,39} may be considered to further improve the estimation performance. However, such priors introduce high computational challenges under ultrahigh-dimensional settings. The feasibility of these developments requires further investigation. Second, in the present study, we employed unsupervised FPCA for feature extraction. Various supervised methods, such as multiscale weighted principal component regression,⁴⁰ may be considered to improve the quality of the extracted features. Third, we proposed a parametric LoI regression model. This parametric framework can be extended to a highly sophisticated semiparametric/nonparametric context. Fourth, to choose the multiple neurocognitive scores that characterize the latent variable in the ADNI study, we utilized a variable “diagnosis” (CN or AD) and conducted two-sample *t*-tests to examine the differences of average scores between CN and AD groups. ADAS13, RAVLT-learning, and FAQ scores were selected based on *P*-values and their popularity in measuring cognitive ability. Our choice may not be optimal, and the use of other scores is worthy of further investigation. One promising approach is to implement an explanatory factor analysis that allows for inclusion of all possible scores

and selects relevant scores based on estimated factor loadings. Finally, the current missing data model does not accommodate the correlation of multiple missing indicators, which is often the case in real applications. A possible solution for addressing this issue is to introduce shared or correlated random effects to the missing data model. We reanalyzed the ADNI data set by introducing a shared random effect to the three missing data models (Equation (8), $j = 1, 2, 3$). The obtained results are similar to those presented in Section 5. Thus, the correlation between the missingness of multiple outcomes does not seem to have a serious impact on estimation in the ADNI study. Nevertheless, this issue is important and worthy of attention in substantive research.

DATA AVAILABILITY STATEMENT

Data sharing not applicable—no new data generated

ORCID

Xinyuan Song  <https://orcid.org/0000-0002-4877-3200>

REFERENCES

1. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
2. Clarke B, Fokoue E, Zhang HH. *Principles and Theory for Data Mining and Machine Learning*. New York, NY: Springer Verlag; 2009.
3. Bühlmann P, Rütimann P, van de Geer S, Zhang CH. Correlated variables in regression: clustering and sparse estimation. *J Stat Plann Infer* 2013;143(11):1835-1858.
4. Goldsmith J, Huang L, Crainiceanu CM. Smooth scalar-on-image regression via spatial Bayesian variable selection. *J Comput Graph Stat*. 2014;23(1):46-64.
5. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *J Am Stat Assoc*. 2006;101(473):119-137.
6. Zhao Y, Ogden RT, Reiss PT. Wavelet-based LASSO in functional linear regression. *J Comput Graph Stat*. 2012;21(3):600-617.
7. Zipunnikov V, Caffo B, Yousem DM, Davatzikos C, Schwartz BS, Crainiceanu C. Functional principal component model for high-dimensional brain imaging. *NeuroImage*. 2011;58(3):772-784.
8. Reiss PT, Ogden RT. Functional generalized linear models with images as predictors. *Biometrics*. 2010;66(1):61-69.
9. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. *J Am Stat Assoc*. 2007;102(479):984-996.
10. Li Y, Wang N, Carroll RJ. Selecting the number of principal components in functional data. *J Am Stat Assoc*. 2013;108(504):1284-1294.
11. Jolliffe IT. A note on the use of principal components in regression. *J Royal Stat Soc Ser C (Appl Stat)*. 1982;31(3):300-303.
12. Hadi AS, Ling RF. Some cautionary notes on the use of principal components regression. *Am Stat*. 1998;52(1):15-19.
13. Ročková V, George EI. The spike-and-slab lasso. *J Am Stat Assoc*. 2018;113(521):431-444.
14. Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J Royal Stat Soc Ser B (Stat Methodol)*. 1999;61(1):173-190.
15. Tang G, Little RJ, Raghunathan TE. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*. 2003;90(4):747-764.
16. Di CZ, Crainiceanu CM, Caffo BS, Punjabi NM. Multilevel functional principal component analysis. *Ann Appl Stat*. 2009;3(1):458.
17. Karhunen K. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae*. 1947;37:1-79.
18. Golub GH, Van Loan CF. *Matrix Computations*, vol. 3. Johns Hopkins University Press: Baltimore, MD; 2013.
19. McCulloch CE, Searle SR, Neuhaus JM. *Generalized Linear, and Mixed Models*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2008.
20. Bollen KA. *Structural Equations with Latent Variables*. New York, NY: Wiley; 1989.
21. Song XY, Lee SY. *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. London, UK: Wiley; 2012.
22. Kang J, Reich BJ, Staicu AM. Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*. 2018;105(1):165-184.
23. d'Aspremont A, Ghaoui LE, Jordan MI, Lanckriet GR. A direct formulation for sparse PCA using semidefinite programming. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT press; 2005:41-48.
24. Ramsay J, Silverman B. *Functional Data Analysis*. 2nd ed. New York, NY: Springer; 2005.
25. Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *J Royal Stat Soc Ser B (Stat Methodol)*. 2009;71(3):671-683.
26. Gelman A, Roberts GO, Gilks WR, et al. Efficient metropolis jumping rules. *Bayesian Stat*. 1996;5:599-607.
27. Jack CR Jr, Bernstein MA, Fox NC, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imag Offic J Int Soc Magn Reson Medic*. 2008;27(4):685-691.
28. McAuliffe MJ, Lalonde FM, McGarry D, Gandler W, Csaky K, Trus BL. Medical image processing, analysis and visualization in clinical research. Paper presented at: Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001;2001:381-386; Bethesda, MD: IEEE.
29. Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*. 2001;13(5):856-876.
30. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp*. 2002;17(3):143-155.

31. Davatzikos C, Genc A, Xu D, Resnick SM. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *NeuroImage*. 2001;14(6):1361-1369.
32. Deluzio K, Astephen J. Biomechanical features of gait waveform data associated with knee osteoarthritis: an application of principal component analysis. *Gait Posture*. 2007;25(1):86-93.
33. Gosche K, Mortimer J, Smith C, Markesbery W, Snowdon D. Hippocampal volume as an index of Alzheimer neuropathology: findings from the Nun Study. *Neurology*. 2002;58(10):1476-1482.
34. Frisoni GB, Prestia A, Rasser PE, Bonetti M, Thompson PM. In vivo mapping of incremental cortical atrophy from incipient to overt Alzheimer's disease. *J Neurol*. 2009;256(6):916.
35. De Jong L, Van der Hiele K, Veer I, et al. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain*. 2008;131(12):3277-3285.
36. Im K, Lee JM, Seo SW, et al. Variations in cortical thickness with dementia severity in Alzheimer's disease. *Neurosci Lett*. 2008;436(2):227-231.
37. Le Carret N, Auriacombe S, Letenneur L, Bergua V, Dartigues JF, Fabrigoule C. Influence of education on the pattern of cognitive deterioration in AD patients: the cognitive reserve hypothesis. *Brain Cogn*. 2005;57(2):120-126.
38. Zhu H, Khondker Z, Lu Z, Ibrahim JG. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *J Am Stat Assoc*. 2014;109(507):977-990.
39. Smith M, Fahrmeir L. Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *J Am Stat Assoc*. 2007;102(478):417-431.
40. Zhu H, Shen D, Peng X, Liu LY. Mwpcr: multiscale weighted principal component regression for high-dimensional prediction. *J Am Stat Assoc*. 2017;112(519):1009-1021.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang X, Song X, Zhu H. Bayesian latent factor on image regression with nonignorable missing data. *Statistics in Medicine*. 2021;40:920–932. <https://doi.org/10.1002/sim.8810>