Scalable Bayesian Image-on-Scalar Regression for Population-Scale Neuroimaging Data Analysis

Yuliang Xu, Timothy D. Johnson, Thomas E. Nichols and Jian Kang*

April 23, 2024

Abstract

Bayesian Image-on-Scalar Regression (ISR) offers significant advantages for neuroimaging data analysis, including flexibility and the ability to quantify uncertainty. However, its application to large-scale imaging datasets, such as found in the UK Biobank, is hindered by the computational demands of traditional posterior computation methods, as well as the challenge of individual-specific brain masks that deviate from the common mask typically used in standard ISR approaches. To address these challenges, we introduce a novel Bayesian ISR model that is scalable and accommodates inconsistent brain masks across subjects in large scale imaging studies. Our model leverages Gaussian process priors and integrates salience area indicators to facilitate ISR. We develop a cutting-edge scalable posterior computation algorithm that employs stochastic gradient Langevin dynamics coupled with memory mapping techniques, ensuring that computation time scales linearly with subsample size and memory usage is constrained only by the batch size. Our approach uniquely enables direct spatial posterior inferences on brain activation regions. The efficacy of our method is demonstrated through simulations and analysis of the UK Biobank task fMRI data, encompassing 8411 subjects and over 120,000 voxels per image, showing that it can achieve a speed increase of 4 to 11 times and enhance statistical power by 8% to 18% compared to traditional Gibbs sampling with zero-imputation in various simulation scenarios.

Keywords: Image-on-Scalar regression; Scalable algorithm; memory-mapping; UK Biobank data; Individual-specific masks.

*To whom correspondence should be addressed: jiankang@umich.edu

1 Introduction

Magnetic Resonance Imaging (MRI) is a non-invasive technique renowned for its comprehensive insight into the brain's structure and function. Functional MRI (fMRI) is an imaging modality that detects neuronal activity via fluctuations in the blood oxygen level dependent (BOLD) signal, providing insights into brain activity (Lindquist, 2008). Task-based fMRI can be used to identify regions where individual traits (e.g. cognitive ability) associate with brain function. We aim to map the influence of a single trait on activity over the whole brain. This represents a classical problem in imaging statistics, where the outcome is an image, and the predictors are multiple scalar variables, commonly known as Image-on-Scalar regression (ISR). As an example, using UK Biobank data (Sudlow et al., 2015), we focus on examining the influence of age on brain activity across the whole brain using task fMRI data. In this scenario, the task fMRI data is the outcome image, while age is the scalar predictor variable.

The analysis of large-scale brain fMRI data presents significant challenges including low signal-to-noise ratio and complex anatomical brain structure (Lindquist, 2008; Smith and Nichols, 2018). The advent of large-scale neuroimaging studies, such as the UK Biobank and Adolescent Brain Cognitive Development (ABCD) study, has introduced new computational challenges for traditional statistical tools in analyzing large-scale fMRI data. These challenges arise from the substantial size of the datasets, as well as the scalability of posterior computation algorithms and the difficulties encountered in achieving convergence in high-dimensional settings.

A specific challenge associated with the UK Biobank fMRI data is the presence of individual-specific brain masks. The brain typically occupies less than half of the cuboid image volume, and a brain mask is used to identify which voxels constitute brain parenchyma and should be included in the analysis. Even after registration of brain images to standard space, each subject's fMRI data can have a unique brain mask, and this is the case in the UK Biobank task fMRI data.

In this paper, we seek to address these challenges by presenting a Bayesian hierarchical model for Image-on-Scalar regression (ISR) that incorporates a sparse and spatially correlated prior on the exposure coefficient. Furthermore, we propose an efficient algorithm utilizing Stochastic Gradient Langevin Dynamics (SGLD)(Welling and Teh, 2011) to ensure scalability. To accommodate the varying individual-specific masks, we employ imputation techniques, enabling us to analyze a wide spatial mask across all individuals.

1.1 UK Biobank Data

The use of imaging biomarkers in clinical diagnostics and disease prognostics has been historically hindered by the lack of imaging data collected before disease onset. The UK Biobank is collecting longitudinal data on one-million UK residents of which a sub-sample of one-hundred thousand are being imaged longitudinally (Miller et al., 2016). The UK Biobank data provides multimodal brain imaging data including structural, diffusion, and functional MRI data. We focus on the task fMRI data, an emotion task where participants are asked to identify faces with negative emotions using shape identification as the baseline task. The objective of this emotional task is to actively involve cognitive functions ranging from sensory and motor areas to regions responsible for processing emotions. Recent studies using the UK Biobank have utilized multimodal data to predict brain age (Cole, 2020), have analyzed the association between resting state connectivity data, education level, and household income (Shen et al., 2018), and have applied deep learning to sex classification in resting state and task fMRI connectomes (Leming and Suckling, 2021); amongst others. (Elliott et al., 2018; Littlejohns et al., 2020).

1.2 Traditional and Recent Practices in ISR

The most common practice for Image-on-Scalar regression is the mass univariate analysis approach (MUA) (Groppe et al., 2011). Although MUA is computationally efficient, easy

to implement, and has well developed multiple comparison correction methods to control false discovery rates, MUA ignores spatial structure and tends to have low statistical power when the data has low signal to noise ratio. To account for spatial dependence, as discussed in Morris (2015), a common approach is to use a low rank approximation. Built on the principle components idea to utilize spatial correlation, Ramsay and Silverman (2005) and Reiss et al. (2010) proposed a penalized regression method using basis expansion to reduce the dimension of the functional outcome. Zhu et al. (2014) propose another model that uses local polynomial and kernel methods for estimating the spatially varying coefficients with discontinuty jumps. Yu et al. (2021) and Li et al. (2021) use bivariate spline functions to estimate the spatial functional estimator supported on the 2-dimensional space. Zhang et al. (2022) propose a quantile regression method that can be applied to high-dimensional outcomes. A common issue with all the aforementioned frequentest methods is that it can be difficult to make inference on the active area selection based on these penalized low rank models. Zhang et al. (2023) develope an efficient deep neural network approach to estimate the spatially varying parameters of very high-dimension with complex structures, but they only focus on point estimation rather than statistical inference. Zeng et al. (2022) propose a Bayesian model with a prior composed of a latent Gaussian variable and a binary selection variable to account for both sparsity and spatial correlation. But the dense covariance matrix can require large memory and computational power when the outcome is very high-dimensional. To address these shortcomings, we propose a scalable Bayesian Image-on-Scalar regression model where the functional coefficient is assigned a sparse and spatially correlated prior, so that we can make inference on the activation areas directly from the posterior inclusion probability (PIP).

1.3 Subject-specific Masks in Brain Imaging

The brain mask used for a multisubject fMRI analysis is typically the intersection of the individual-specific masks, only analyzing voxels were all subjects have data. However, in the UK Biobank, an intersection mask of 8,411 subjects reduces the analysis volume by 35% relative to average subject mask volume. Some authors will attempt to minimise this effect by identifying subjects with particularly small masks, but this is a laborious process and discards data. For a particular subject, all voxels that lie in the set difference of the union mask and that subject's mask results in missing values for said subject. Many of the voxels with missing data occur around the edge of the union mask (Mulugeta et al., 2017). Historically, researchers usually do not account for missing data, however, for PET data, missing values are imputed using a *soft mean* (Hammers et al., 2007) derived using available data at each voxel.

In volumetric fMRI data, individual brain masks can be slightly different from one another, due to some subject's brain falling outside of the field-of-view truncation, residual variation in inter-subject brain shape not accounted for by atlas registration, and susceptibility-induced signal loss. Especially, voxels above the nasal cavity and the above the ear canals suffer from signal loss and are classified as non-brain tissue in a subject-specific manner.

All brain image analyses requires a brain mask to avoid wasted computation and uninterpretable results on non-brain voxels. For mass-univariate fMRI analyses, some authors created custom software (Szaflarski et al., 2006; Maullin-Sapey and Nichols, 2022) or used mixed effect models with longitudinal fMRI data on each voxel (Szaflarski et al., 2012) to account for subject-specific masks, or explicitly accounted for missing data with multiple imputation (Vaden Jr et al., 2012). Among the major fMRI software packages, however, neither FSL nor SPM can account for subject-specific masks, and only AFNI's 3dMEMA (Chen et al., 2012) (simple group analysis) or 3dLME (Chen et al., 2013) (general mixed effects) accounts for subject-specific masks. However, all of these methods are mass-univariate.

To the best of our knowledge, our work is the first spatial Bayesian method to account for varying missing data patterns over the brain. By using individual masks with imputation, we make full use of all collected data.

1.4 Scalable Posterior Algorithms

The scalable posterior algorithm we use is based on the stochastic gradient Langevin dynamics (SGLD) algorithm (Welling and Teh, 2011). The SGLD algorithm is effective at handling large-scale data as it approximates the posterior gradient using subsamples of the data. Variations on the SGLD algorithm for scalable posterior sampling have been proposed. Wu et al. (2022) proposed a Metropolis-Hasting algorithm using mini-batches where both the proposal and acceptance probability are both approximated by the current mini-batch. Kim et al. (2022) proposed an adaptive SGLD algorithm (Adam SGLD) that sets a preconditioner for SGLD and allows the gradient at different directions to update with different step sizes. Aside from these MCMC algorithms, variational Bayesian inference (Jaakkola and Jordan, 1999) is another popular option for approximating the posterior mean in high-dimensional settings. There has been increasing use of variational inference for high-dimensional posteriors such as imaging data analysis (Kaden et al., 2008; Kulkarni et al., 2022), and various scalable extensions of variational inference (Hoffman et al., 2013; Ranganath et al., 2014; Blaiotta et al., 2016). Although these variational meethods are computationally efficiency, our goal, however, is to obtain the entire MCMC sample that provides uncertainty quantification for the activation areas of interest. Hence, we settled on the SGLD type algorithm for its scalability.

The main contributions of our proposed method are

1. to provide an efficient posterior computation algorithm for Bayesian Image-on-Scalar regression, scalable to large sample size and high resolution image;

2. to introduce the individual-specific brain masks and expand the analysis region from

an intersection mask of all individuals to an inclusive mask using imputation.

In particular, our method uses batch updates and memory-mapping techniques to analyze large sample imaging data that is too big to fit into random access memory on many computers. In addition, we provide an imputation-based method that allows us to handle individual-specific masks and makes full use of the observed data.

The rest of this paper is organized as follows. In Section 2, we introduce our Bayesian Image-on-Scalar model. In Section 3, we discuss algorithm and computational details. In Section 4, we demonstrate the performance of our proposed algorithms against the classical MUA approach under different simulation settings. In Section 5, we apply out proposed method to a subset of the UK Biobank imaging data and provide sensitivity analysis results. In Section 6 we summarize our contribution and provide a general discussion. Additional simulation results and real data sensitivity analysis can be found in the Supplementary Materials. Our implementation is provided as an R package, SBIOS, and is publicly available on Github [1].

2 Model

Let $\mathcal{B} \subset \mathbb{R}^3$ denote the entire brain region. Let $\{s_j\}_{j=1}^p \subset \mathcal{B}$ be a set of fixed grid points in $\mathcal{B}$, on which we observe brain image intensity values. For individual i ($i = 1, \ldots, n$), let $Y_i(s_j)$ be the image intensity at voxel s_j. To incorporate the individual-specific masks, let $\mathcal{V}_i$ denote the set of locations where the image intensity is observed for individual i, i.e., for any $s_j \in \mathcal{V}_i$, $Y_i(s_j)$ is not missing. For any $i = 1, \ldots, n$, $\mathcal{V}_i \subset \{s_1, \ldots, s_p\}$. Let X_i be the primary covariate of interest, Z_{ik} be the k-th confounding covariate for $k = 1, \ldots, q$. We propose an

[1] See the SBIOS R package https://github.com/yuliangxu/SBIOS

Image-on-Scalar regression model. For individual i and any $s_j \in \mathcal{V}_i$,

$$Y_i(s_j) = X_i\beta(s_j)\delta(s_j) + \sum_{k=1}^{m} \gamma_k(s_j)Z_{ik} + \eta_i(s_j) + \epsilon_i(s_j), \quad \epsilon_i(s_j) \sim \mathrm{N}(0, \sigma_Y^2). \tag{1}$$

The spatially varying parameter $\beta(s)$ estimates the magnitude in the image intensity that can be explained by the predictor X, and the binary selection indicator $\delta(s)$ follows a Bernoulli prior with selection probability $p(s)$. In practice, we set $p(s) = 0.5$ for any $s \in \mathcal{B}$ as the prior for $\delta(s)$. The selection variable $\delta(s)$ determines the active voxels in the brain associated with the predictor X and $\beta(s)\delta(s)$ is of main interest. The spatially varying parameter $\gamma_k(s)$ is the coefficient for the k-th confounder Z_k, and $\eta_i(s)$ accounts for individual level spatially correlated noise. By introducing the individual effect η_i as a parameter, we are separating spatially correlated noise from spatially independent noise ϵ_i and we can safely assume a completely independent noise term $\epsilon_i(s)$ across all locations s_j, hence avoiding large-scale covariance matrix computations in the noise term. This is similar to the correlated noise model in Zhu et al. (2014). Proposition 1 in Zhang et al. (2023) provides sufficient conditions for model identifiability (1): 1) the design matrix $\tilde{\mathbf{X}} := (\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{n\times(m+1)}$ is a full rank matrix, and 2) for any i and any $s \in \mathcal{B}$, denote $\boldsymbol{\eta}(s) = (\eta_1(s), \ldots, \eta_n(s)) \in \mathbb{R}^n$, $\tilde{\mathbf{X}}^T\boldsymbol{\eta}(s) = 0$.

In addition, we define the missing rate for individual i as $\tilde{m}_i = p^{-1}\sum_{j=1}^{p}\{1 - I(s_j \in \mathcal{V}_i)\}$ where $I(s_j \in \mathcal{V}_i)$ is an indicator function taking value 1 if individual i has observed data at location s_i and 0 otherwise.

For model (1), we specify the following prior distributions for the following paramaters:

$$\delta(s) \sim \text{Ber}\{p(s)\}, \quad \text{for any } s \in \mathcal{B} \tag{2}$$

$$\beta(s) \sim \mathcal{GP}(0, \sigma_\beta^2 \kappa), \quad \text{for any } s \in \mathcal{B} \tag{3}$$

$$\gamma_k(s) \sim \mathcal{GP}(0, \sigma_\gamma^2 \kappa), \quad \text{for any } s \in \mathcal{B}, \quad j = 1, \ldots, q \tag{4}$$

$$\eta_i(s) \sim \mathcal{GP}(0, \sigma_\eta^2 \kappa), \quad \text{for any } s \in \mathcal{B} \tag{5}$$

The spatially-varying functional coefficients $\beta(s), \gamma_k(s), \eta_i(s)$ are assumed to have Gaussian Process (GP) priors with mean 0 and kernel function $\sigma^2\kappa(\cdot,\cdot)$, where σ^2 can be different for each functional parameter. Popular choices of the kernel function κ including the exponential square kernel and the Matérn kernel (8). We use the latter in both simulation studies and real data analysis as it offers flexible choices of the kernel parameters. Here, all GPs are assumed to have the same kernel function κ for computational efficiency. In addition, we expect the individual effects η_i to be smoother than β, hence choosing an appropriate kernel for β is sufficient for the estimation of η_i. In the UK Biobank data analysis, κ is chosen to reflect the outcome image $Y_i(s)$ correlation structure in a data adaptive way, as discussed in Section 5.

3 Posterior Computation

3.1 Posterior Sampling with Gaussian Process Priors

The 3D task fMRI image data is divided into R regions using the Harvard-Oxford cortical and subcortical structural atlases (Desikan et al., 2006). Between-region independence is assumed when constructing the Gaussian kernel for β, γ_k and η_i. Instead of assuming a whole brain correlation structure for the GP priors, a block diagonal covariance structure is computationally efficient and allows us to capture more detailed information within each

region, especially regions of smaller size and complex spatial structures.

To sample from the posterior of the GPs, we use a basis decomposition approach. By Mercer's theorem (Rasmussen and Williams, 2005), for any $\beta(s) \sim \mathcal{GP}(0, \sigma_\beta^2 \kappa)$, we can use a basis decomposition,

$$\beta(s) = \sum_{l=1}^{\infty} \theta_{\beta,l} \phi_l(s), \quad \theta_{\beta,l} \sim \mathcal{N}(0, \sigma_\beta^2 \lambda_l),$$

where λ_l is the l-th eigenvalue, and ϕ_l is the l-th eigenfunction (see Section 4.2 in Rasmussen and Williams (2005)). The eigenvalues in this expansion satisfy $\sum_{l=1}^{\infty} \lambda_l < \infty$, and the eigenfunctions form an orthonormal basis in $L^2(\mathcal{B})$, i.e. $\int_{s \in \mathcal{B}} \phi_l(s) \phi_{l'}(s) ds = I(l = l')$, where $I(\cdot)$ is an indicator function taking value 1 if the expression inside the bracket is true. Hence, we use the coefficient space of $\theta_{\beta,l}$ rather than $\beta(s)$. Similarly, we expand $\gamma_k(s) = \sum_{l=1}^{\infty} \theta_\beta \phi_l(s)$ and $\eta_i(s) = \sum_{l=1}^{\infty} \theta_{\eta,i,l} \phi_i(s)$, where $\theta_{\gamma,k,l}$ and $\theta_{\eta,i,l}$ are the basis coefficients for γ_k and η_i respectively. In practice, only a finite number $l = 1, \ldots, L$ of $\{\theta_{\beta,l}\}_{l=1}^{\infty}$ are used (corresponding to the L largest eigenvalues), and $\beta(s)$ is approximated by $\sum_{l=1}^{L} \theta_{\beta,l} \phi_l(s)$. This basis decomposition approach is applied for all three GP priors (3) - (5).

With the basis expansion coefficient, we can sample from the L-dimensional space to model the p-dimensional image data. We also partition the brain into regions to further speed up computation and assume a region-independence structure for the spatially varying parameters β, γ_k, and η_i. Assume there are $r = 1, \ldots, R$ regions that form a partition of the mask $\mathcal{B}$, denoted as $\mathcal{B}_1, \ldots, \mathcal{B}_R$. For the three GP priors (3) - (5), we assume that the kernel function $\kappa(s_j, s_k) = 0$ for any $s_j \in \mathcal{B}_r, s_k \in \mathcal{B}_{r'}, r \neq r'$, and the prior covariance matrix on the fixed grid has a block diagonal structure.

For the r-th region, let p_r be the number of voxels in $\mathcal{B}_r$ and let $Q_r = (\psi_l(s_{r,j}))_{l=1,j=1}^{L_r, p_r} \in \mathbb{R}^{L_r \times p_r}$ be the matrix with the (l, j)-th component $\psi_l(s_{r,j})$ where $\{s_{r,j}\}_{j=1}^{p_r}$ forms the fixed grid in $\mathcal{B}_r$. Because of the basis approximation with cutoff L_r, Q_r is not necessarily an orthornormal matrix, hence we use the QR decomposition to get an approximately orthornormal Q_r,

i.e. $Q_r^T Q_r = I_{L_r}$, where I_{L_r} is the identity matrix. With the region partition, the GP priors on the r-th region can be reexpressed as $\beta_r = (\beta(s_{r,1}), \ldots, \beta(s_{r,p_r}))^T \approx Q_r \boldsymbol{\theta}_{\beta,r}$, where $\boldsymbol{\theta}_{\beta,r} \sim \mathcal{N}\left(0, \sigma_\beta^2 D_r\right)$ and D_r is a diagonal matrix with diagonal $(\lambda_{r,1}, \ldots, \lambda_{r,L_r}) \in \mathbb{R}^{L_r}$.

To present the working model with region partitions, denote $Y_{i,r}^* = Q_r^T Y_{i,r} \in \mathbb{R}^{L_r}$ as the low-dimensional mapping of the ith image on the r-th region where $Y_{i,r} = \{Y_i(s_j)\}_{s_j \in \mathcal{B}_r} \in \mathbb{R}^{p_r}$, and $\epsilon_{i,r}^* = Q_r^T \epsilon_{i,r}$, $\epsilon_{i,r} = \{\epsilon_i(s_j)\}_{s_j \in \mathcal{B}_r} \in \mathbb{R}^{p_r}$. Let $\text{diag}\{x\}$ be the diagonal matrix with diagonal x. Let $\boldsymbol{\delta}_r = \{\delta(s_j)\}_{s_j \in \mathcal{B}_r} \in \mathbb{R}^{p_r}$. After basis decomposition

$$Y_{i,r}^* = Q_r^T X_i \text{diag}\{\boldsymbol{\delta}_r\} Q_r \boldsymbol{\theta}_{\beta,r} + \sum_{k=1}^{m} \theta_{\gamma,j,r} Z_j + \theta_{\eta,i,r} + \epsilon_{i,r}^* \tag{6}$$

with the prior specification $\boldsymbol{\theta}_{\beta,r} \sim N(0, \sigma_\beta^2 D_r)$, $\boldsymbol{\theta}_{\gamma,k,r} \sim N(0, \sigma_\gamma^2 D_r)$, $\boldsymbol{\theta}_{\eta,i,r} \sim N(0, \sigma_\eta^2 D_r)$, $\epsilon_{i,r}^* \sim N(0, \sigma_Y^2 I_{L,r})$. The working model (6) based on regionally independent kernels performs a whole brain analysis since $\sigma_\beta^2, \sigma_\gamma^2, \sigma_\eta^2$, and σ_Y^2 are estimated globally across regions. This is also the first step towards reducing memory cost by using a low-dimensional approximation. The finite cutoff L is chosen to reflect the flexibility of the true functional parameters: fewer bases are required to approximate the smooth function $\beta(s)$. In our real data analysis, L is determined by extracting the eigenvalues of the covariance kernel matrix. For one brain region, first compute the $p \times p$ dimensional covariance matrix with appropriately tuned covariance parameters, get the eigenvalues of such covariance matrix, and choose the cutoff such that the summation $\sum_{l=1}^{L} \lambda_l$ is over 90% of $\sum_{l=1}^{p} \lambda_l$. Details for choosing the covariance parameters in (8) and sensitivity analysis on the 90% cutoff are discussed in Sections 5 and 5.4.

3.2 Scalable Algorithm for Large Dataset

Inspired by Welling and Teh (2011) and Wu et al. (2022), we propose Algorithm 2 based on stochastic gradient Langevin dynamics (SGLD) to sample the posteriors from a large data

set. The SGLD algorithm outperforms Gibbs sampling (GS) in three ways: 1) under the assumption that all individuals are independently and identically distributed according to the proposed model, the SGLD algorithm allows us to compute the log likelihood on a small subset of data, making it computational efficient compared to GS; 2) the SGLD algorithm adds a small amount of noise to the gradient of the log posterior density, making it more effective in exploring the posterior parameter space; and 3) as the step size decreases to 0, the SGLD algorithm provides a smooth transition from the stochastic optimization stage to the posterior sampling stage.

We apply the SGLD algorithm on θ_β. Denote $\theta_\beta^{(t)}$ as the value at the t-th iteration. Let τ_t be the step size at the t-th iteration, let $\pi(\theta_{\beta,l})$ denote the prior, and let $\pi_{i\in\mathcal{I}}\left(Y_i \mid X_i, Z_i, \theta\right) := \prod_{i\in\mathcal{I}} \pi\left(Y_i \mid X_i, Z_i, \theta\right)$ be the likelihood for a subsample $\mathcal{I}$, where θ is the collection of all parameters. Denote by n_s the number of subjects in the subsample $\mathcal{I}$. Let $\nabla f(x)$ be the gradient of $f(x)$. At the t-th iteration, the l-th component in θ_β is updated as

$$\theta_{\beta,l}^{(t)} \longleftarrow \theta_{\beta,l}^{(t-1)} + \frac{\tau_t}{2}\nabla \log \pi(\theta_{\beta,l}^{(t-1)}) + \frac{\tau_t}{2}\frac{n}{n_s}\nabla \log \pi_{i\in\mathcal{I}}\left(Y_i \mid X_i, Z_i, \theta^{(t-1)}\right) + \sqrt{\tau_t}\varepsilon_l \tag{7}$$

where $\varepsilon_l \overset{\text{iid}}{\sim} N(0,1)$. Let L_r be the number of basis coefficients for the r-th brain region. The time complexity for (7) is $\min\{O(L_r^3), O(L_r^2 n_s)\}$. The full sample size n is usually significantly larger than L_r, hence approximating the full likelihood with a subsample of the data significantly decreases the computational complexity.

Based on the mini-batch idea of the SGLD algorithm, we propose the following Algorithm 2, where the large data set is first split into B smaller batches. Each batch of data is loaded using memory-mapping techniques. Within each batch, a small subsample of size n_s is randomly drawn to be used at each iteration. By splitting the full data into B batches, we reduce the auxiliary space complexity for computing the required summary statistics down to the size of B, instead of the size of the full data.

On line 10 of Algorithm 2, using Gibbs sampling to update γ, δ also requires the entire

Algorithm 1 Scalable Bayesian Image-on-Scalar (SBIOS) regression with memory mapping

1: Set subsample size s, an integer t_I for the frequency to update η_i, and initial values for all parameters.
2: Split the entire sample into B batches, sequentially load each batch of data and save each batch to the disk using memory-mapping. Set batch index b to 1.
3: **for** iteration $t = 1, 2, \ldots, T$ **do**
4: Update the step size τ_t for t-th iteration.
5: Load batch b into memory.
6: **for** region $r = 1, ..., R$ **do**
7: Randomly select a subsample $\mathcal{I}$ of size n_s from batch b.
8: Update $\theta_{\beta,l}$ for region r using (7) based on the selected subsample.
9: **end for**
10: Update $\gamma, \delta, \sigma_\gamma, \sigma_\beta$ using Gibbs sampling.
11: $b = b + 1$. If $b > B$, set $b = 1$.
12: **if** t is a multiple of t_I **then**
13: Iterate through all batches to update $\{\eta_i\}_{i=1}^N$.
14: Update σ_Y, σ_η using Gibbs sampling.
15: (Optional) Impute the missing outcome $Y_i(s_j)$ for all the missing voxel indices $s_j \notin \mathcal{V}_i$.
16: **end if**
17: **end for**

data set, but the posterior distributions of γ and δ, in fact, only rely on summary statistics that can be pre-computed based on the entire data set. In practice, at the beginning of the algorithm, we iterate through each batch once to compute these summary statistics and directly use them to update γ and δ at each iteration. The same cannot be done for$\theta_{\beta,l}$, because $\delta(s)$ can be 0 or 1, and the posterior variance for $(\theta_{\beta,1}, \ldots, \theta_{\beta,L})^T$ depends on $Q^T \text{diag}\{\delta\} Q$, which is no longer a diagonal matrix and requires updating at each iteration. See Equation (6). Hence, sampling $\theta_{\beta,l}$ is more computationally demanding than sampling $\theta_{\gamma,l}$. We provide detailed derivations of the posterior distribution for $\theta_{\beta,l}, \theta_{\gamma,l}, \delta(s)$ in the Supplementary Material, Section 1.

On line 13 of Algorithm 2, since $\theta_{\eta,i,l}$ is the coefficient for individual-level effect, we must iterate through all samples to update all of the $\theta_{\eta,i,l}$. As such, storing $\theta_{\eta,i,l}$ in memory grows as the sample size n increases. As $\eta_i(s)$ more of a nuisance parameter and not ourmain focus, we choose to update η_i less frequently. Further improvements to memory allocation can be

implemented if we also use batch-splitting on η and save the samples of η as file-backed matrices. This is implemented for the `SBIOS0` function in our package as an example.

For the optional imputation found on line 15 of Algorithm 2, imputing $Y_i(s_j)$ where $s_j \notin \mathcal{V}_i$ means that the missing values in Y_i and all summary statistics associated with Y_i need to be updated every t_I iterations. Hence we use an index-based updating scheme. For each i, denote the complementary set $\mathcal{V}_i^c := \{s_j\}_{j=1}^p - \mathcal{V}_i$ as the index set of all missing voxels for individual i. We keep track of $\mathcal{V}_i^c$, and create a vector `Y_imp` to store the imputed outcome values only on those indices in $\mathcal{V}_i^c$ for each i, and update the corresponding summary statistics every time $Y_i(s_j), s_j \in \mathcal{V}_i^c$ is updated. The time complexity to update all missing values in Y_i is $O(L \times (|\mathcal{V}_i^c|))$ where $|\mathcal{V}_i^c|$ is the number of missing voxels for individual i, and L is the total number of basis coefficients. We provide a detailed algorithm for updating missing outcomes in the Suppmentary Materials.

Algorithm 2 is implemented using the Rcpp package `bigmemory` (Kane et al., 2013). The `bigmemory` package allows us to store large matrices on disk as a `big.matrix` class and extract the address of the large matrices. In the beginning of Algorithm 2, when the entire data is split into smaller batches, each batch is loaded in R as a `big.matrix` class, then the address of these big matrices are passed to Algorithm 2, accessing different batches of data becomes very efficient and memory-conserving.

4 Simulations

4.1 Simulation Design

We compare our proposed method with existing methods via a simulation studies. We consider the following four different methods.

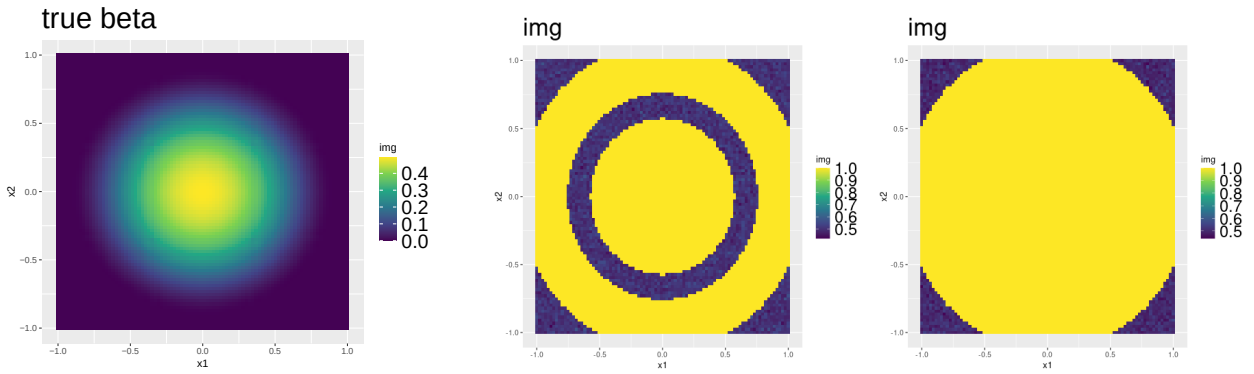
Mass Univariate Analysis (MUA) is one of the most commonly used method for ISR models. MUA ignores any spatial correlation in the image and treats the data at all voxels as

completely independent and performs linear regression on the data at each voxel. Ignoring spatial correlation may lead to potential overfitting. When comparing selection accuracy for $\beta(s)\delta(s)$, we use the Benjamini-Hochberg adjusted p-values (Benjamini and Hochberg, 1995) to control the false discovery rate. The missing voxels are directly imputed as 0 in the MUA.

Bayesian Image-on-Scalar (BSIOS) model is the baseline Bayesian method for our proposed model (1)-(5), where all parameters are sampled via Gibbs sampling. Missing voxels are imputed as 0. The sampling algorithm reads the entire data set into memory.

Scalable Bayesian Image-on-Scalar model with 0 imputation (SBIOS0) is our proposed model (1)-(5) with sampling perfomed via the SGLD algorithm and memory-mapping. Missing data at voxels in individual masks are be replaced by 0.

Scalable Bayesian Image-on-Scalar model with GS-imputation (SBIOSimp) is the proposed model (1)-(5) using the SBIOS0 algorithm, however, missing data in the individual masks are imputed from the proposed model.



(a) True covariate $\beta(s)\delta(s)$. (b) Missing Pattern I (c) Missing Pattern II

Figure 1: True signal and missing patterns. Missing Pattern I & II: Inclusion rate ranges from 0.5 to 1.

To demonstrate the scalability of the proposed method, three simulation studies are used to show the selection accuracy (Simulation I), maximum memory usage (Simulation II), and time scalability (Simulation III). In each simulation, the entire sample is split into B batches of data, each containing 500 subjects, with subsample size n_s fixed at 200. The true image $\beta(s)$ is a 2D image of size $p = 60 \times 60 = 8100$, as shown in Figure 1 (a). The binary variable

$\delta(s)$ is generated as 1 when $\beta(s)$ is nonzero, and 0 otherwise. The confounder parameters $\gamma_k(s)$ and individual effects $\eta_i(s)$ are generated based on the coefficients $\theta_{\gamma,k,l}, \theta_{\eta,i,l}$, sampled from the standard normal distribution. Figure 1b and 1c are visual illustrations of the spatial missing rate over all samples, where the image value at each pixel s_j is $n^{-1}\sum_{i=1}^{n} I(s_j \in \mathcal{V}_i)$. The difference between these 2 missing patterns is that missing pattern I selects part of the active area where $\beta(s) \neq 0$ as missing at a rate of 50% while missing pattern II only selects missing voxels outside of the active area.

When generating the covariance kernel function, the 90×90 2D images are evenly split into 9 (3×3) regions that are assumed to be independent from each other, a priori. Hence we compute the basis decomposition of each region supported on a square of 30×30 pixels, where the kernel function is the Matérn kernel with $\rho = 2, \nu = 1/5$ as in (8).

$$\kappa(s', s; \nu, \rho) = C_\nu(\|s' - s\|_2^2/\rho),\ C_\nu(d) := \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}d\right)^{\nu} K_\nu(\sqrt{2\nu}d), \tag{8}$$

where K_ν is a modified Bessel function of the second kind (Rasmussen and Williams, 2005). The number of basis coefficients used is 10% of the number of pixels, hence we have 90 basis coefficients for each region. For a less smooth input signal pattern, we can use a larger number of basis functions to increase the flexibility of the fitted functional coefficients. In the real data application, we use a data adaptive method to choose the number of bases, and provide a sensitivity analysis on the number of bases selected (Sections 5 and 5.4).

All three Bayesian methods (BIOS, SBIOS0, SBIOSimp) use the same basis decomposition of the kernel function in (8), and all run for a total of 5000 MCMC iterations, where the last 1000 iterations are used for posterior inference. The prior for $\delta(s)$ is Bernoulli(0.5). For the SGLD based methods (SBIOS0, SBIOSimp), we tune the step size τ_t across iterations using the formula $\tau_t = a(b+t)^{-\gamma}$ as in Welling and Teh (2011), and set $a = 0.001, b = 10, \gamma = 0.55$ to ensure that through the 5000 MCMC iterations, the step size s_t decreases from around 2×10^{-3} to 9×10^{-6}. The initial values for all three Bayesian methods iden-

tical, where $\delta(s)$, $\theta_{\beta,l}$ and $\theta_{\gamma,k,l}$ are set to 1, $\theta_{\eta,i,l}$ are set to 0, and all variance parameters $\sigma_Y, \sigma_\beta, \sigma_\gamma, \sigma_\eta$ are set to 1.

4.1.1 Evaluation Criteria

To assess the variable selection accuracy, we compute the True Positive Rate (TPR) when the False Positive Rate (FPR) is controlled at 10%. Because of the selection variable $\delta(s)$, we can obtain a Posterior Inclusion Probability (PIP) from the $m = 1, \ldots, M$ posterior MCMC samples of $\delta(s)$,

$$\text{PIP}(s) = \frac{1}{M} \sum_{m=1}^{M} \delta_m(s).$$

Setting a threshold between 0 and 1 on $\text{PIP}(s)$ gives a mapping of the active pixels. Hence we choose 20 evenly-spaced points between 0 and 1 to fit an ROC curve with linear splines, and get the estimated TPR when FPR is at 10% from the fitted ROC curve. For MUA, we choose the cutoff on p-values to be the 20 quantiles of the Benjamini-Hochberg (BH) adjusted p-values corresponding to the probabilities at the 20 evenly-space points, and use the same method to obtain TPR when FPR controlled at 10%.

All three Bayesian methods (BIOS, SBIOS0, SBIOSimp) are implemented using Rcpp (Eddelbuettel and François, 2011) with RcppArmadillo (Eddelbuettel and Sanderson, 2014). Code for all simulations and all implementations of the proposed methods and competing methods can be found in the R package SBIOS on Github.

4.2 Simulation Results

We demonstrate the advantage of the proposed SBIOS algorithm with respect to three objectives: variable selection accuracy, maximum memory usage, and scalability in terms of computing time as the sample size increases. The main results for selection accuracy, memory usage, and time scalability are shown in Figures 2, 4, and 5a respectively.

4.2.1 Simulation I: Variable Selection

In simulation I, to compare variable selection accuracy, when generating the individual masks, we provide results under 2 different patterns of the masks as shown in Figure 1 (b) and (c). Figure 1 provides an illustration where in each pattern there exists a common area, and each location inside of this common area has observed data for all individuals. Outside of the common area we allow missingness. This setting is motivated by the UK Biobank fMRI images where most missing voxels are on the edge of brain regions, and there is a large common area in the center of the brain with fully observed data. Outside of the commonly observed area, we allow missing rate at each voxel $n^{-1}\sum_{i=1}^{n} I(s_j \in \mathcal{V}_i^c)$ to be 0.5 or 0.9. The simulation I is performed under a total of 8 combinations of different scenarios: (i) missing percentage is 0.5 or 0.9; (ii) $\sigma_Y = 0.5$ or $\sigma_Y = 1$; (iii) missing pattern follows Figure 1 (b) or (c). The different missing percentages and missing patterns show the advantage of using imputation method (SBIOSimp), and the different σ_Y represent performances under different signal-to-noise ratios.

The selection accuracy result in Figure 2 shows that with Missing Pattern I, with FPR controlled at 10%, SBIOSimp has much better TPR across different settings compared to the other methods. Within the 0-imputation based methods, the Bayesian methods (BIOS, SBIOS0) have similar performances, and are generally better than MUA, except for the case where $\sigma_Y = 0.5$, missing percentage is 0.5. In situations where estimating $\beta\delta$ becomes challenging, such as when the signal-to-noise ratio is low or the missing percentage is high, spatially-correlated methods can offer an advantage.

With Missing Pattern II, the three Bayesian methods outperform MUA in all settings. But since Missing Pattern II only generate missing pixels on non-active region, there is smaller difference in FPR-controlled TPR between SBIOSimp and other methods, as shown in the Supplementary Material Section 2.

Comparing four sub-figures in Figure 2, we see that (i) as the missing percentage increases

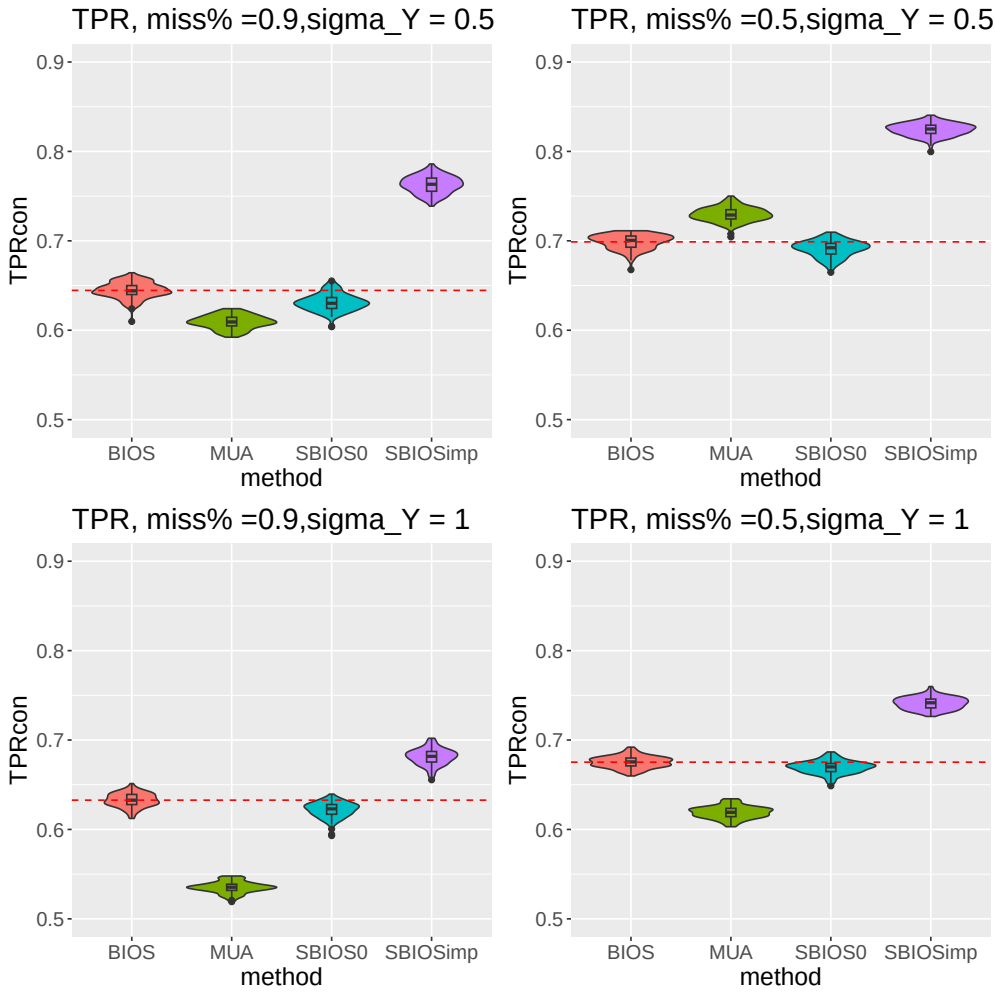


Figure 2: True Positive Rate based on 100 replicated simulations when False Positive Rate is controlled at 0.1 when $n = 3000, p = 8100$. Missing Pattern I.

(comparing right subplots to left subplots), TPR decreases for all methods, with MUA most sensitive to the change in missing percentage; (ii) as the signal to noise ratio decreases (σ_Y goes from 0.5 to 1, comparing top subplots to bottom subplots), TPR decreases, also with the biggest gap in MUA.

Since in real data analysis, we cannot directly tune FPR according to the true signal, we can use Morris FDR control method (Meyer et al., 2015) to select the cutoff on PIP when determining the active region. The three Bayesian methods are tuned using Morris FDR control method, and MUA still uses BH correction. The target FDR is set to be 0.05, and the threshold for the BH-adjusted p-value is set to be 0.05. Figure 10 shows the true FDR

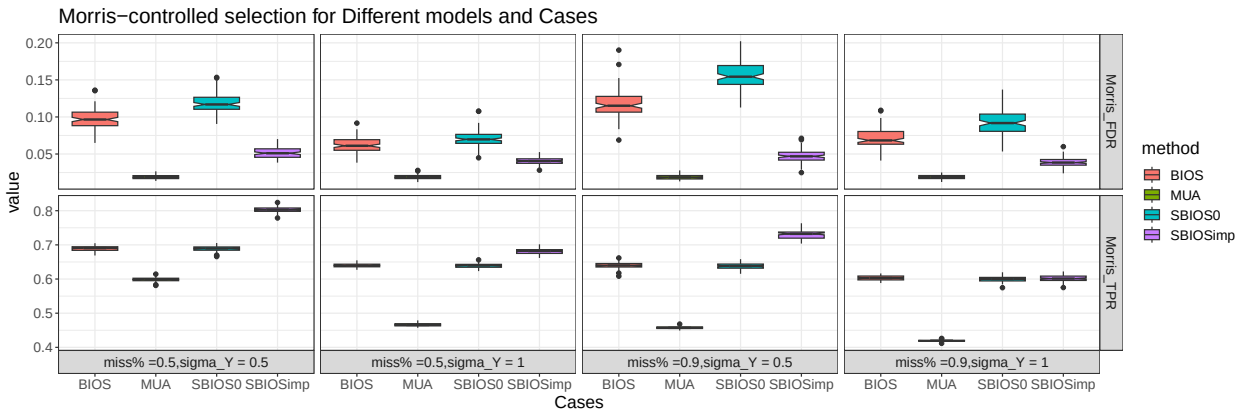


Figure 3: True Positive Rate and False Discovery Rate using Morris FDR control method, based on 100 replicated simulations, $n = 3000, p = 8100$. Missing Pattern I.

and TPR for all settings using this FDR control method. SBIOSimp overall has the best performance with FDR comparable to MUA, and the highest TPR. But we do observe that the Morris FDR control method cannot necessarily achieve the target FDR in practice.

4.2.2 Simulation II: Memory Usage

In simulation II, to compare the memory usage, the generative setting uses the case $\sigma_Y = 1$, missing pattern II, and the missing rate is 0.5. The sample size $n = 3000, 6000, 9000, 12000$, with fixed $p = 3600$. This simulation is only to test the memory efficiency, SBIOSimp has similar memory usage to SBIOS0 except for a vector to save the imputed outcome, so we are only comparing MUA, BIOS and SBIOS0.

To record the maximum memory usage, we use the R function `Rprof`. In particular, we use `summaryRprof` with the option `memory="tseries"` where the memory usage is recorded for segments of time intervals, and we take the maximum memory across all such segments as the maximum memory usage. The reported memory usage is for running each algorithm alone, not including the memory usage for reading the data into memory. The reading memory is separately recorded from the running memory, and the detailed results can be found in the Supplementary Material Section 2. For MUA and BIOS, we directly read all data into memory first, and pass the address to BIOS. For SBIOS0 and SBIOSimp, each

batch of data are read into R as file-backed matrices on the disk, and its address is passed to both SBIOS0 and SBIOSimp for sampling.

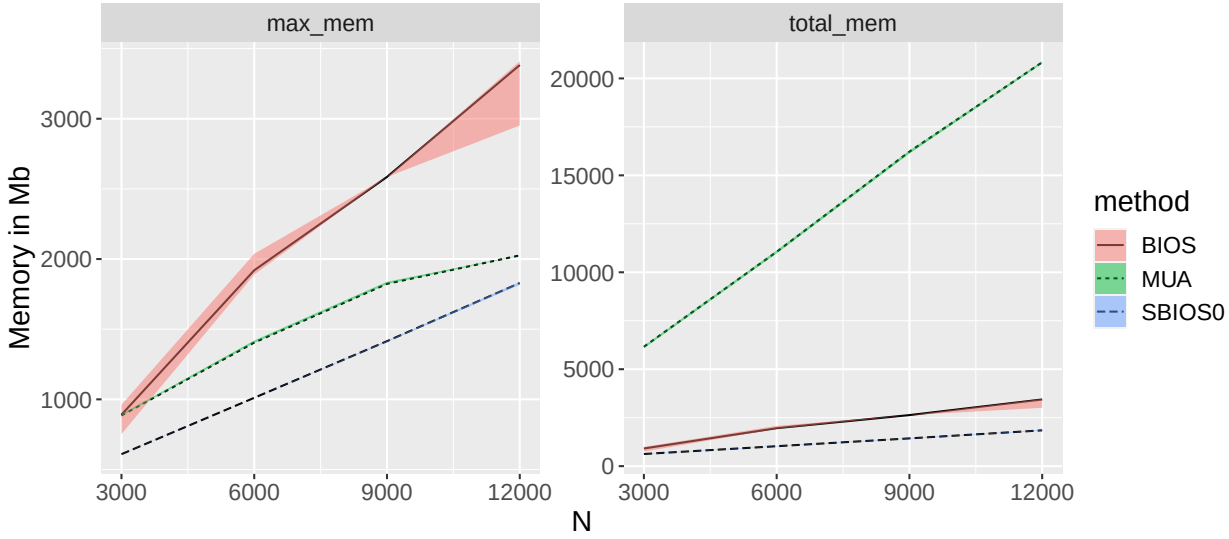
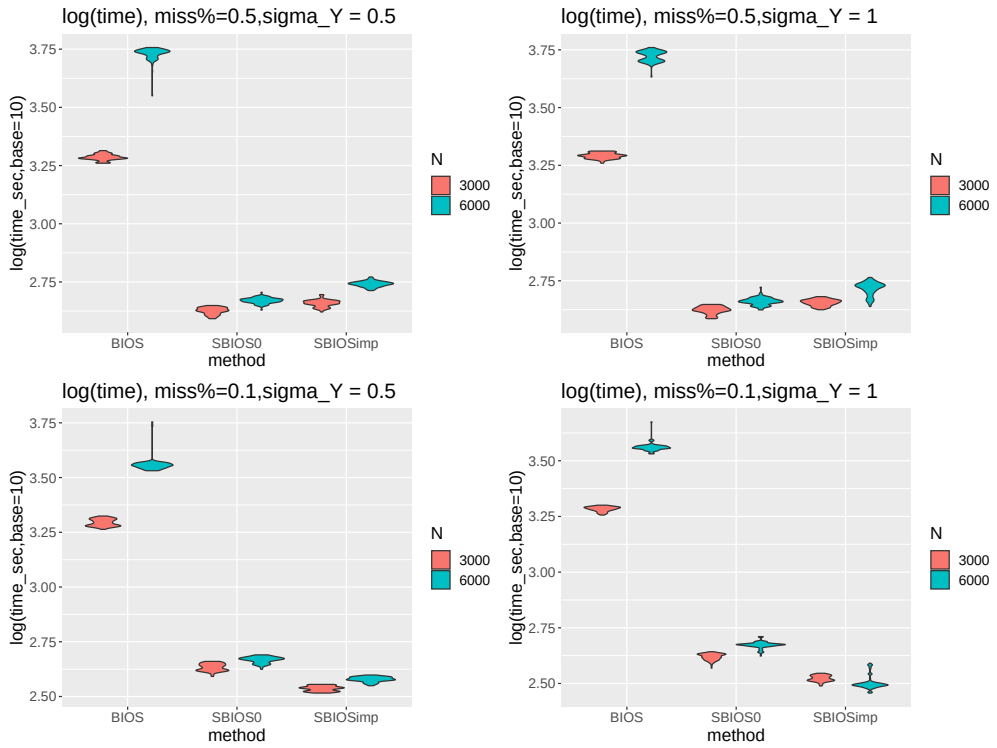


Figure 4: Memory cost in Mb over $n = 3000, 6000, 9000, 12000$, with 100 replications. The left panel is the maximum memory usage, and the right panel is the total memory usage. The mean, 97.5% and 2.5% quantile lines are plotted.

The result is shown in Figure 4. The left panel in Figure 4 reflects the peak memory usage when running each method, and is the main result of interest. In terms of maximum memory usage over time, our proposed method SBIOS0 is comparable to MUA, which only loops through simple linear regressions, and is much more efficient compared to the standard Bayesian method BIOS. In terms of total memory usage over time (right panel in Figure 4), both Bayesian methods take much less total memory compared to MUA which needs to loop through p locations in total, but our proposed method SBIOS0 still outperforms the other.

4.2.3 Simulation III: Time Scalability

In simulation III, we compare the time scalability for BIOS, SBIOS0, and SBIOSimp, and exclude MUA result since MUA only needs to run p simple linear regressions and is instantaneous. For this simulation, the generative setting uses the case with missing pattern II, and the sample size $n = 3000, 6000$, with fixed $p = 3600$.



(a) Computing time for Bayesian methods, as seconds in the log scale of base 10, for both $n = 3000$ and $n = 6000$ based on 100 replicated simulation results.

method	SBIOS0		SBIOSimp	
n	3000	6000	3000	6000
miss%=0.5, $\sigma_Y = 0.5$	4.57	11.50	4.24	9.77
miss%=0.1, $\sigma_Y = 0.5$	4.58	7.78	5.73	9.56
miss%=0.5, $\sigma_Y = 1$	4.66	11.53	4.32	10.11
miss%=0.1, $\sigma_Y = 1$	4.61	7.72	5.76	11.44

(b) Ratio of speed for SBIOS0 and SBIOSimp when compared to BIOS, i.e. time of BIOS over time of SBIOS0 (or SBIOSimp) under 4 cases and n = 3000 and 6000, based on 100 replicated simulation results.

Figure 5: Computational time comparison

As illustrated in Figure 5a, the computing time (seconds) is reported in the log scale of base 10. Our proposed SBIOS0 and SBIOSimp are shown to have much better time scalability across all settings compared to BIOS that only uses Gibbs sampling.

5 UK Biobank Application

In this real data application, we use 3D task fMRI data from UK Biobank as the outcome, age as the exposure variable. The three confounding variables are gender, the interaction between age and gender, and head size.

5.1 Data Preprocessing and Estimation Procedure

The outcome data we use is the emotion task fMRI data, where participants are asked which of two faces with fearful expressions (or shapes) presented on the bottom of the screen match the faces (or shapes) at the top of the screen. (See details in the Hariri faces/shapes emotion task (Miller et al., 2016; Hariri et al., 2002), as implemented in the Human Connectome Project (HCP).) The dimension of the 3D fMRI data in MNI atlas space is a cube of $91 \times 109 \times 91$ voxels, and we used a random selection of $n = 8,411$ subjects with task fMRI. The NIFTI data is about 27 GB, and the processed RDS files are about 8 GB. The original NIFTI outcome data for one subject is around 3Mb, and the NIFTI mask data of binary format for one subject is around 26Kb. Using R package `RNifti` (Clayden et al., 2023), the time taken for preprocessing 1 subject of data, including directly loading the outcome and mask data from DropBox folder, is around 2.3 seconds, and saving the preprocessed data into file-backed matrices is instantaneous. We use the Harvard-Oxford cortical and subcortical structural atlases (Desikan et al., 2006). After preprocessing, we have a total of 110 brain regions.

To define the analysis mask, we first define the inclusion rate at location s_j to be $h(s_j) = n^{-1} \sum_{i=1}^{n} I(s_j \in \mathcal{V}_i)$, where $\mathcal{V}_i$ is the set of all observed locations for individual i. We define the group analysis mask as $\mathcal{B} = \{s_j : h(s_j) > 0.5\}$, i.e., the area where each voxel has at least 50% observed data. As shown in Figure 6, compared to $\mathcal{B}$ where the inclusion rate is at least 50% (blue area), a vast area is missing from the group analysis mask (100% inclusion rate; purple area overlaying on top of the blue area), especially the bottom of the brain,

including the orbitofrontal cortex and inferior temporal cortex, and notably the amygdala region important in this emotion task. In particular, the atlas we are using contains 110 brain regions, but the mask with 100% inclusion rate contains only 101 regions.

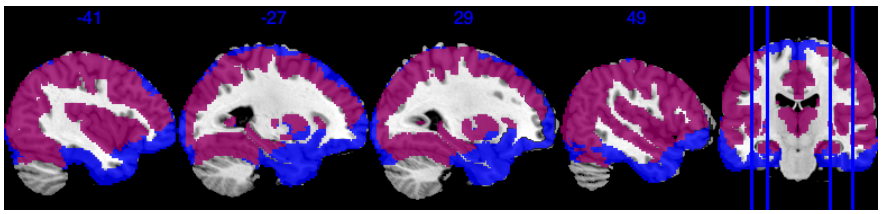


Figure 6: Analysis mask using an inclusion rate threshold of 50% and an intersection mask (100% inclusion rate). Purple area indicates 100% inclusion; blue area indicates only the mask of inclusion rate above 50% but less than 100%.

After applying the 50% inclusion rate common mask $\mathcal{B}$, we end up with $\mathcal{B}$ that contains a total number of $p = 122,049$ voxels. The image outcome $Y_i(s)$ are standardized across subjects, and the continuous covariates are also standardized. For each region, we apply Matérn kernel function but with different ρ and ν parameters (8), to account for the different smoothness of each region. The ρ and ν are tuned so that the empirical covariance of $Y(s_j), Y(s'_j)$ would roughly match the estimated covariance by the Matérn kernel function given ρ and ν. The number of basis L is chosen such that the cumulative summation $\sum_{l=1}^{L} \lambda_l$ accounts for 90% of the total summation of all eigenvalues, hence we have a total number of $L = 26,235$. In Section 5.4, we provide a sensitivity analysis when the cutoff is based on 92% of the total summation.

The total of $n = 8,411$ subjects are split into 9 batches of data, with each batch containing 950 or 811 subjects. The subsample size is set to 200, and the step size decaying parameters $a = 0.0001, b = 1, \gamma = 0.35$ are chosen so that the step size would roughly decrease from 7×10^{-5} to 5×10^{-6} over a total of 5000 MCMC iterations. To speed up the convergence, we first run 5000 iterations with η fixed at 0, use the posterior mean of the last 1000 iterations as the initial values for β and γ, and then use these initial values of β and γ to run the SBIOS0 and SBIOSimp methods. The initial values for $\theta_{\eta,i,l}$ are set as 0 everywhere, and the initial

values for $\delta(s)$ are set as 1 everywhere. The initial values for $\sigma_Y, \sigma_\eta, \sigma_\beta, \sigma_\gamma$ are all set to be 0.1. To check the convergence of SBIOSimp, we run 4 chains and use Gelman and Rubin test. The l_2 norm of the residual is computed for the last 1000 iterations from 4 chains, and the point estimate of the Gelman-Rubin statistic is 1.18 with an upper confidence limit at 1.5, indicating that the MCMC chain has approximately converged.

5.2 Analysis Result

We present top 10 regions identified by the SBIOSimp method in Table 1. With the posterior sample for each region, we can compute the Region Level Activation Rate (RLAR) as follows: denote $\mathcal{B}_j$ as the mask for region j, the RLAR is $\sum_{s\in\mathcal{B}_j}\delta(s)/|\mathcal{B}_j|$, where $|\mathcal{B}_j|$ is the total number of voxels in region j. Hence for each MCMC sample of $\delta(s)$, we can obtain one sample of RLAR for all regions, and therefore obtain the posterior distribution of RLAR for all regions. We present histograms of the posterior distribution of the RLAR over the last 1000 MCMC iterations in Supplementary Material Seciton 3. Table 1 presents the top 10 regions with the highest posterior mean of RLAR and their 95% credible intervals. To present the marginal effect of each $\beta(s_j)$, we also report the effect summing over each region computed separately for the positive and negative effect. Only voxels with a marginal posterior inclusion probability (PIP) over 95% are viewed as active voxels. Because the top 10 regions has no active voxels with positive effects, we only report the negative effect in Table 1. We also report the number of active voxels in Table 1. All active voxels in the top 10 regions have negative effect, which indicates that as age increases, the cognitive ability to process tasks tends to decrease. This trend can also be reflected in the raw data, where we provide scatter plots of age against the average image intensity in the Supplementary Material Section 3. To interpret the numeric result in Table 1, use the *Right temporal fusiform cortex, anterior division* region as an example. There are on average 98% of locations with brain activities that are associated with age within this region, and for 1 standard deviation

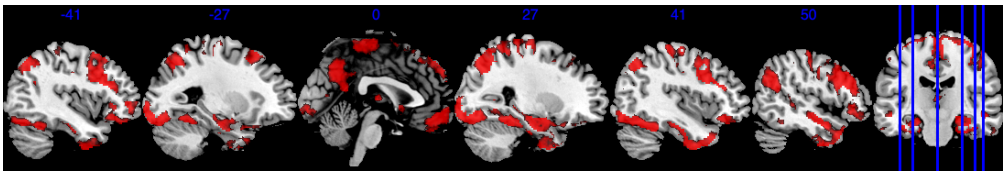
increase in the standardized age, there will be a decrease of 16.79 of brain signal intensity summed over all locations with in this region.

Table 1: Top 10 regions ordered by Region Level Activation Rate (RLAR), and their 95% credible interval. For the region names, (R) means right hemisphere, (L) means left hemisphere, AD means anterior division, PT means pars triangularis. The last columns report the negative voxel effect size summed over each region i.e. $\sum_{j\in\mathcal{B}_j}\mathbb{E}\{\beta(s_j)|\beta(s_j)<0\}$, and inside the bracket are the number of negative voxels. Only voxels with marginal inclusion probability greater than 0.95 are included, otherwise counted as zero effect voxel. Hence positive voxels are omitted due to low inclusion probability.

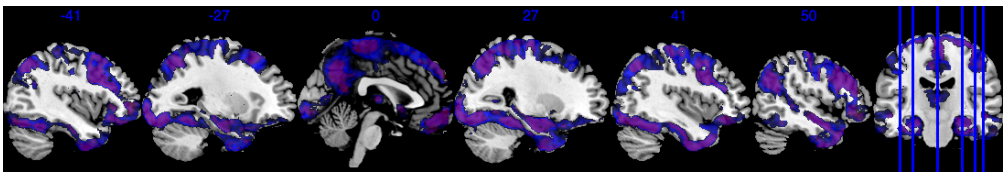
Region Name	Size	RLAR	Neg Sum (Count)
(R) temporal fusiform cortex, AD	276	0.98 (0.97,1.00)	-16.79 (260)
(L) inferior frontal gyrus, PT	692	0.98 (0.97,0.99)	-41.51 (626)
(R) inferior temporal gyrus, AD	314	0.97 (0.96,0.99)	-18.36 (266)
(L) temporal fusiform cortex, AD	301	0.96 (0.94,0.98)	-15.35 (225)
(R) amygdala	215	0.96 (0.94,0.98)	-14.47 (190)
(R) hippocampus	296	0.95 (0.93,0.97)	-13.22 (226)
(L) inferior temporal gyrus, AD	348	0.94 (0.92,0.97)	-14.58 (227)
(R) middle temporal gyrus, AD	402	0.94 (0.92,0.96)	-17.65 (290)
(R) frontal medial cortex	365	0.93 (0.91,0.96)	-15.36 (256)
(L) frontal medial cortex	444	0.92 (0.90,0.94)	-16.24 (261)

For a visual representation, Figure 7a shows the voxel level PIP in the brain sagittal view. The highlighted red region represents voxels with higher than 0.95 PIP. Figure 7b presents the effect size of $\beta(s)\delta(s)$, with the highlighted area in the range between -0.06 and -0.03. The overlaying purple area in Figure 7b is the same as the red area in Figure 7a, in order to demonstrate that voxels with PIP greater than 0.95 also correspond to voxels with larger absolute value of effect size. We notice that the active area (defined by voxel level PIP greater than 0.95) has a negative effect $\beta(s)$. This can also be validated by the scatter plot in Supplementary Material Section 3, where the image outcome generally has a negative association with age across all individuals. In Supplementary Material Section 3, we also provide the Morris FDR control result for both the SBIOS0 and SBIOSimp methods. Because Morris FDR control method tends to over-select the active voxels, we do not use this as the main reference.

Based on our results, we have the following general interpretations: (i) when controlling



(a) Posterior inclusion probability (PIP) thresholded in $[0.95, 1]$. Sagittal view. The first two sagittal slices are in the left side of the brain.



(b) Posterior mean of $\beta(s)\delta(s)$, thresholded in $[-0.06, -0.03]$. The overlaying purple area is the mask of PIP greater than 0.95.

Figure 7: Illustration of result on a grayscale brain background image (`ch2bet`, Holmes et al. (1998)). Images are created using MRIcron (Rorden and Brett, 2000).

for the confounders, age has a negative impact on the neural activity for emotion-related tasks; (ii) the negative effect reflected from each voxel is of very small scale, shown as in Figure 7b, indicating a very low voxel level signal-to-noise ratio; (iii) the top 3 brain regions with the highest RLAR are (a) *Right temporal fusiform cortex, anterior division*, considered as a key structure for face perception, object recognition, and reading (Weiner and Zilles, 2016), (b) *Left inferior frontal gyrus, pars triangularis*, an area for semantic processing of language (Foundas et al., 1996), and (c) *right inferior temporal gyrus, anterior division*, an area for language and semantic memory processing, visual perception, and multimodal sensory integration (Onitsuka et al., 2004). These top 3 regions are also consistently identified in the sensitivity analysis when using half of the data as training data, see Section 5.4.

5.3 Comparing SBIOSimp with SBIOS0

To check the differences between SBIOS0 and SBIOSimp, we compare the posterior mean of $\beta(s)$ given $\text{PIP}(s) \geq 0.95$ between SBIOS0 and SBIOSimp with different levels of the inclusion rates on these regions in Figure 8. Each point in Figure 8 represents the effect of age on the brain signal on one location in the brain. The 6 regions in Figure 8 are chosen from

the top 10 regions in Table 1 with high missingness. Comparing blue dots (low inclusion rate, $h(s_j) \in [0.5, 0.7)$), red dots (medium inclusion rate, $h(s_j) \in [0.7, 0.9)$) with black dots (high inclusion rate, $h(s_j) \in [0.9, 1]$), we can see that β fitted with SBIOS0 on voxels with lower inclusion rate tend to be closer to 0 or directly mapped to 0 according to $I(\text{PIP}(s) \geq 0.95)$, compared with SBIOSimp. This implies that by directly imputing missing outcomes with 0, SBIOS0 tends to put more shrinkage on the posterior mean of $\beta(s)$ compared to SBIOSimp. Hence SBIOS0 potentially has lower power than SBIOSimp to detect the signals, which could be justified by the simulation result shown in Figure 2.

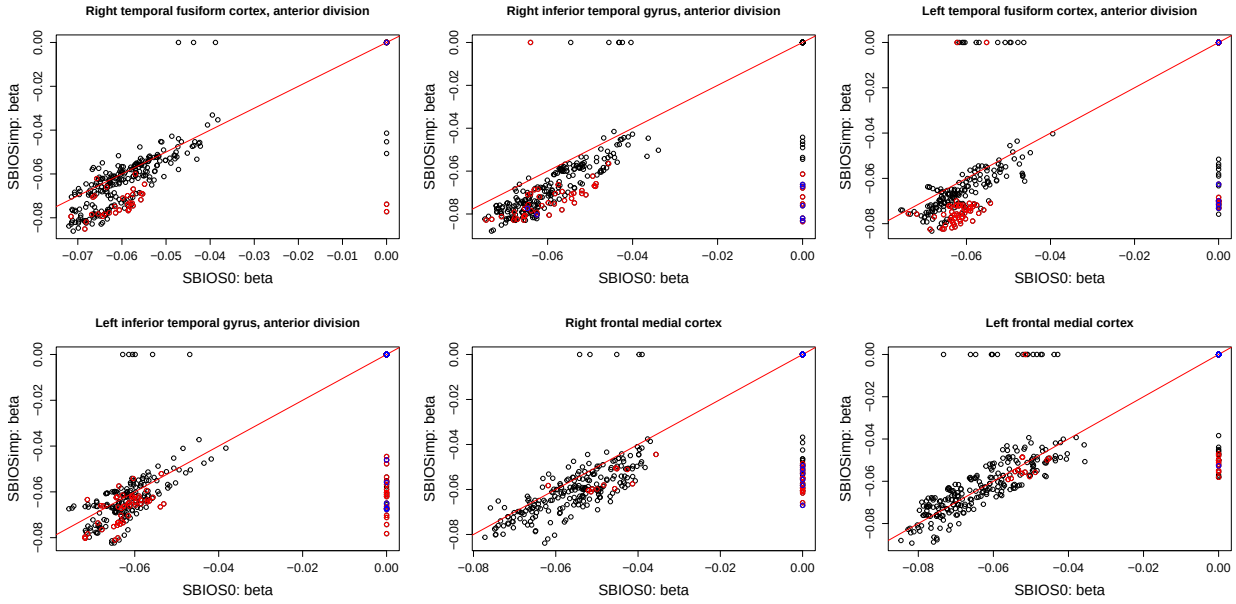


Figure 8: Scatter plot of the posterior mean of $\beta(s_j)I(\text{PIP}(s_j) \geq 0.95)$ based on SBIOS0 and SBIOSimp on six selected regions with high missingness. Blue dots indicate voxels with inclusion rate $h(s_j) \in [0.5, 0.7)$. Red dots indicate voxels with inclusion rate $h(s_j) \in [0.7, 0.9)$. Black dots indicate voxels with inclusion rate $h(s_j) \in [0.9, 1]$.

5.4 Sensitivity Analysis

To further validate the result reported in subsection 5.2, we conduct sensitivity analysis on different choices of the hyperparameter $IG(a, b)$ in the prior for σ_Y^2, the choice of σ_β^2, and the number of basis in the Gaussian kernel. The baseline setting for results in subsection

5.2 is $\sigma_Y^2 \sim IG(0.1, 0.1)$, hyperparameter $\sigma_\beta^2 = 0.01$, and the number of basis is based on 90% of total eigenvalues ($L = 26235$). In the sensitivity analysis, in case 1, the prior for σ_Y^2 is $IG(1, 1)$, and $IG(0.1, 0.1)$ for case 2. In case 3, we use $\sigma_\beta^2 = 1$. In case 4, we choose the number of basis based on 92% of total eigenvalues ($L = 29541$). Because Case 1 & 2 have very similar results, we report them together.

First, we examine the impact of sensitivity parameters on the training and testing Root Mean Square Erorr (RMSE) by splitting the data into training (N=4701) and testing (N=3710) data sets, and use SBIOSimp to fit the training data. The training and testing RMSE are shown in Table 2b. The choice of these parameters have very little effect on the RMSE. Next, we use the four cases of sensitivity parameters to fit the full data, and report the RLAR and its rank among all 110 regions.

Table 2a presents the region level difference between each sensitivity cases and the final result presented in subsection 5.2. Cases 1 and 2 produce almost the same result and are reported together. We can see that the top three regions, right temporal fusiform cortex in anterior division, left inferior frontal gyrus in pars triangularis, right inferior temporal gyrus in anterior division, are consistently selected as the top three regions with highest region level active rate. The top 10 regions with highest RLAR selected by case 1 to 4 are the same, but with slightly different rankings.

In the Supplementary Material, we also provide the top 10 region ranked by the active ratio, a voxel level marginal measure, which is defined as the ratio of active voxels in each region. The result is also consistent among the 4 cases, with slightly different ranking.

6 Conclusion and Discussion

We propose a Bayesian hierarchical model for Image-on-Scalar regression, with computationally efficient algorithm for posterior sampling, and apply our proposed method on UK Biobank application. Our proposed model is able to capture high dimensional spatial correla-

Table 2: Sensitivity analysis of SBIOSimp on UK Biobank data.

(a) Sensitivity Analysis Results. The RLAR for each region is reported, with the rank (decreasingly ordered by region level IP) of each region inside the brackets. For the region names, (R) means right hemisphere, (L) means left hemisphere, AD means anterior division, PT means pars triangularis.

Region ID	Region Name	Cases 1&2	Case 3	Case 4
37	(R) temporal fusiform cortex, AD	0.99(1)	0.99(1)	0.99(1)
53	(L) inferior frontal gyrus, PT	0.98(2)	0.98(2)	0.99(2)
14	(R) inferior temporal gyrus, AD	0.97(3)	0.97(3)	0.98(3)
85	(L) temporal fusiform cortex, AD	0.96(4)	0.95(5)	0.97(4)
109	(R) amygdala	0.96(5)	0.96(4)	0.96(5)
108	(R) hippocampus	0.95(6)	0.95(6)	0.95(7)
62	(L) inferior temporal gyrus, AD	0.94(7)	0.94(7)	0.96(6)
11	(R) middle temporal gyrus, AD	0.94(8)	0.93(8)	0.95(8)
25	(R) frontal medial cortex	0.93(9)	0.92(9)	0.94(9)
73	(L) frontal medial cortex	0.92(10)	0.91(10)	0.93(10)

(b) Training and testing RMSE for the 4 sensitivity analysis cases. RMSE is the root mean square error across all subjects and all voxels.

	Cases 1&2	Case 3	Case 4
Settings	$\sigma_Y^2 \sim IG(1,1)$, $\sigma_Y^2 \sim IG(0.1, 0.1)$	$\sigma_\beta^2 = 1$	92% basis, L = 29541
train RMSE	68.39879	68.39739	68.39701
test RMSE	60.82462	60.82442	60.82389

tion, account for the individual level correlated noise, and provide uncertainty quantification on the active area selection through posterior inclusion probability. Our main computational contribution is to use a scalable SGLD algorithm for big data ISR, to use memory-mapping technique to solve the memory limit for large-scale imaging data, and to utilize individual specific masks using imputation approach to make use of as much imaging data as possible.

Through extensive simulation studies, we compare different implementation of the proposed model, from the standard Gibbs sampling implementation (BIOS), the SGLD implementation with missing outcome imputed as 0 (SBIOS0), and the proposed imputation based SGLD algorithm (SBIOSimp), with the standard Mass Univariate Analysis (MUA). The simulation results provide evidence that SBIOSimp can achieve better variable selection accuracy, with much smaller memory cost, and the computational time is scalable to large-scale data. In the UK Biobank application, we also validate our results by providing

sensitivity analysis under different hyper-parameter and kernel settings.

Although the performance on simulations and UK Biobank data indicates great potential in applying our method to other large-scale imaging applications, there are still several limitations. First of all, the underlying Gaussian process requires a user-defined kernel function. Currently, our data-adaptive way of choosing kernel parameters is still partially subject to the user's discretion. Secondly, we assume that the individual effect η share the same kernel with β and γ mainly because of computational efficiency, but in practice there might be individual level latent confounders than cannot be captured by estimating η. Another limitation comes from SGLD algorithm. Although the main focus is to identify active areas and effect size of β, and SGLD can give us accurate first moment estimation of β. However, as the step size decreases to zero, the posterior MCMC sample may not truly recover the variance of β. A future direction of this work is to apply the proposed method to other large scale imaging data such as the ABCD study Casey et al. (2018).

References

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, 57, 289–300.

Blaiotta, C., Cardoso, M. J., and Ashburner, J. (2016), "Variational inference for medical image segmentation," *Computer Vision and Image Understanding*, 151, 14–28.

Casey, B., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., et al. (2018), "The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites," *Developmental cognitive neuroscience*, 32, 43–54.

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., and Cox, R. W. (2013), "Linear mixed-effects modeling approach to FMRI group analysis," *Neuroimage*, 73, 176–190.

Chen, G., Saad, Z. S., Nath, A. R., Beauchamp, M. S., and Cox, R. W. (2012), "FMRI group analysis combining effect estimates and their variances," *Neuroimage*, 60, 747–765.

Clayden, J., Cox, B., and Jenkinson, M. (2023), *RNifti: Fast R and C++ Access to NIfTI Images*, r package version 1.4.5.

Cole, J. H. (2020), "Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors," *Neurobiology of aging*, 92, 34–42.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006), "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *Neuroimage*, 31, 968–980.

Eddelbuettel, D. and François, R. (2011), "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40, 1–18.

Eddelbuettel, D. and Sanderson, C. (2014), "RcppArmadillo: Accelerating R with high-performance C++ linear algebra," *Computational Statistics and Data Analysis*, 71, 1054–1063.

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., Marchini, J., and Smith, S. M. (2018), "Genome-wide association studies of brain imaging phenotypes in UK Biobank," *Nature*, 562, 210–216.

Foundas, A. L., Leonard, C. M., Gilmore, R. L., Fennell, E. B., and Heilman, K. M. (1996), "Pars triangularis asymmetry and language dominance." *Proceedings of the National Academy of Sciences*, 93, 719–722.

Groppe, D. M., Urbach, T. P., and Kutas, M. (2011), "Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review," *Psychophysiology*, 48, 1711–1725.

Hammers, A., Asselin, M.-C., Hinz, R., Kitchen, I., Brooks, D. J., Duncan, J. S., and Koepp, M. J. (2007), "Upregulation of opioid receptor binding following spontaneous epileptic seizures," *Brain*, 130, 1009–1016.

Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., and Weinberger, D. R. (2002), "The amygdala response to emotional stimuli: a comparison of faces and scenes," *Neuroimage*, 17, 317–323.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013), "Stochastic variational inference," *Journal of Machine Learning Research*.

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., and Evans, A. C. (1998), "Enhancement of MR images using registration for signal averaging," *Journal of computer assisted tomography*, 22, 324–333.

Jaakkola, T. S. and Jordan, M. I. (1999), "Variational probabilistic inference and the QMR-DT network," *Journal of artificial intelligence research*, 10, 291–322.

Kaden, E., Anwander, A., and Knösche, T. R. (2008), "Variational inference of the fiber orientation density using diffusion MR imaging," *Neuroimage*, 42, 1366–1380.

Kane, M. J., Emerson, J., and Weston, S. (2013), "Scalable Strategies for Computing with Massive Data," *Journal of Statistical Software*, 55, 1–19.

Kim, S., Song, Q., and Liang, F. (2022), "Stochastic gradient Langevin dynamics with adaptive drifts," *Journal of statistical computation and simulation*, 92, 318–336.

Kulkarni, P. H., Merchant, S., and Awate, S. P. (2022), "Mixed-dictionary models and variational inference in task fMRI for shorter scans and better image quality," *Medical Image Analysis*, 78, 102392.

Leming, M. and Suckling, J. (2021), "Deep learning for sex classification in resting-state and task functional brain networks from the UK Biobank," *NeuroImage*, 241, 118409.

Li, X., Wang, L., Wang, H. J., and Initiative, A. D. N. (2021), "Sparse learning and structure identification for ultrahigh-dimensional image-on-scalar regression," *Journal of the American Statistical Association*, 116, 1994–2008.

Lindquist, M. A. (2008), "The Statistical Analysis of fMRI Data," *SSO Schweiz. Monatsschr. Zahnheilkd.*, 23, 439–464.

Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J. D., Boultwood, C., Collins, R., Conroy, M. C., et al. (2020), "The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions," *Nature communications*, 11, 2624.

Maullin-Sapey, T. and Nichols, T. E. (2022), "BLMM: Parallelised computing for big linear mixed models," *NeuroImage*, 264, 119729.

Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P., and Morris, J. S. (2015), "Bayesian function-on-function regression for multilevel functional data," *Biometrics*, 71, 563–574.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., et al. (2016), "Multimodal population brain imaging in the UK Biobank prospective epidemiological study," *Nature neuroscience*, 19, 1523–1536.

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

Mulugeta, G., Eckert, M. A., Vaden, K. I., Johnson, T. D., and Lawson, A. B. (2017), "Methods for the analysis of missing data in FMRI studies," *Journal of biometrics & biostatistics*, 8.

Onitsuka, T., Shenton, M. E., Salisbury, D. F., Dickey, C. C., Kasai, K., Toner, S. K., Frumin, M., Kikinis, R., Jolesz, F. A., and McCarley, R. W. (2004), "Middle and inferior temporal gyrus gray matter volume abnormalities in chronic schizophrenia: an MRI study," *American Journal of Psychiatry*, 161, 1603–1611.

Ramsay, J. O. and Silverman, B. W. (2005), *Fitting differential equations to functional data: Principal differential analysis*, Springer.

Ranganath, R., Gerrish, S., and Blei, D. (2014), "Black Box Variational Inference," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, eds. Kaski, S. and Corander, J., Reykjavik, Iceland: PMLR, vol. 33 of *Proceedings of Machine Learning Research*, pp. 814–822.

Rasmussen, C. E. and Williams, C. K. I. (2005), *Gaussian processes for machine learning*, Adaptive Computation and Machine Learning series, London, England: MIT Press.

Reiss, P. T., Huang, L., and Mennes, M. (2010), "Fast function-on-scalar regression with penalized basis expansions," *The international journal of biostatistics*, 6.

Rorden, C. and Brett, M. (2000), "Stereotaxic display of brain lesions," *Behavioural neurology*, 12, 191–200.

Shen, X., Cox, S. R., Adams, M. J., Howard, D. M., Lawrie, S. M., Ritchie, S. J., Bastin, M. E., Deary, I. J., McIntosh, A. M., and Whalley, H. C. (2018), "Resting-state connectivity and its association with cognitive performance, educational attainment, and household income in the UK Biobank," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3, 878–886.

Smith, S. M. and Nichols, T. E. (2018), "Statistical challenges in "big data" human neuroimaging," *Neuron*, 97, 263–268.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015), "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, 12, e1001779.

Szaflarski, J. P., Altaye, M., Rajagopal, A., Eaton, K., Meng, X., Plante, E., and Holland, S. K. (2012), "A 10-year longitudinal fMRI study of narrative comprehension in children and adolescents," *Neuroimage*, 63, 1188–1195.

Szaflarski, J. P., Schmithorst, V. J., Altaye, M., Byars, A. W., Ret, J., Plante, E., and Holland, S. K. (2006), "A longitudinal functional magnetic resonance imaging study of language development in children 5 to 11 years old," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 59, 796–807.

Vaden Jr, K. I., Gebregziabher, M., Kuchinsky, S. E., and Eckert, M. A. (2012), "Multiple imputation of missing fMRI data in whole brain analysis," *Neuroimage*, 60, 1843–1855.

Weiner, K. S. and Zilles, K. (2016), "The anatomical and functional specialization of the fusiform gyrus," *Neuropsychologia*, 83, 48–62.

Welling, M. and Teh, Y. W. (2011), "Bayesian learning via stochastic gradient Langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, Citeseer, pp. 681–688.

Wu, T.-Y., Rachel Wang, Y., and Wong, W. H. (2022), "Mini-batch metropolis–hastings with reversible SGLD proposal," *Journal of the American Statistical Association*, 117, 386–394.

Yu, S., Wang, G., Wang, L., and Yang, L. (2021), "Multivariate spline estimation and inference for image-on-scalar regression," *Statistica Sinica*, 31, 1463–1487.

Zeng, Z., Li, M., and Vannucci, M. (2022), "Bayesian Image-on-Scalar Regression with a Spatial Global-Local Spike-and-Slab Prior," *Bayesian Analysis*, 1, 1–26.

Zhang, D., Li, L., Sripada, C., and Kang, J. (2023), "Image response regression via deep neural networks," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkad073.

Zhang, Z., Wang, X., Kong, L., and Zhu, H. (2022), "High-dimensional spatial quantile function-on-scalar regression," *Journal of the American Statistical Association*, 117, 1563–1578.

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.

A Posterior derivation

We provide the posterior distributions for each parameter without considering individual masks first, and then provide the derivation for imputing the missing outcomes when considering individual masks.

We use I_d to denote the identity matrix in $\mathbb{R}^{d\times d}$. We use $\text{diag}\{D\}$ to denote a diagonal matrix whose diagnal vector is D.

A.1 Posterior densities assuming $Y(s)$ is fully observed for $\forall s \in \mathcal{B}$

Consider the basis expansion approximations of the GPs

$$\beta(s) \approx \sum_{l=1}^{L} \theta_{\beta,l}\psi_l(s) \tag{9}$$

$$\gamma_k(s) \approx \sum_{l=1}^{L} \theta_{\gamma,k,l}\psi_l(s) \tag{10}$$

$$\eta_i(s) \approx \sum_{l=1}^{L} \theta_{\eta,il}\psi_l(s) \tag{11}$$

where $\theta_{\beta,l} \overset{ind}{\sim} N(0, \sigma_\beta^2\lambda_l)$. Denote $D = (\lambda_1, \ldots, \lambda_L)^T$ as the vector of eigenvalues.

Let $Q \in \mathbb{R}^{p\times L}$ be the basis decomposition matrix, where $Q_{j,l} = \psi_l(s_j)$. Here, to approximate the orthonormality of $\psi_l(s)$, $Q^TQ = I_L$, Q is made an orthonormal matrix using QR decomposition.

Denote $Y_i^* = Q^TY_i \in \mathbb{R}^L$ as the low-dimensional mapping of the ith image $Y_i \in \mathbb{R}^p$. After basis decomposition

$$Y_i^* = Q^TX_i\text{diag}\{\delta\}Q\theta_\beta + \sum_{k=1}^{m} \theta_{\gamma,j}Z_{i,k} + \theta_{\eta,i} + \epsilon_i^* \tag{12}$$

$$\theta_\beta(s) \sim N(0, \sigma_\beta^2\text{diag}\{D\}) \tag{13}$$

$$\theta_{\gamma,k}(s) \sim N(0, \sigma_\gamma^2\text{diag}\{D\}) \tag{14}$$

$$\epsilon_i^*(s) \sim N(0, I_L) \tag{15}$$

The posterior distribution can be derived as follows.

$$\Sigma_{\beta(s)}|\text{rest} = \left(\frac{1}{\sigma_\beta^2}\text{diag}\left\{D\right\}^{-1} + \sum_{i=1}^{n} Q^T\text{diag}\left\{X_i\delta\right\}Q\left(\frac{1}{\sigma_Y^2}I\right)Q^T\text{diag}\left\{X_i\delta(s)\right\}Q\right)^{-1}$$

$$\theta_\beta(s)|\text{rest} \sim N\left(\Sigma_{\theta_\beta(s)|\text{rest}}\left\{\frac{1}{\sigma_Y^2}\sum_{i=1}^{n}\left\{[Y_i^* - \theta_{\eta,i} - \sum_{k=1}^{m}\theta_{\gamma,k}Z_{i,k}]^T Q^T\text{diag}\{X_i\delta(s)\}Q\right\}\right\}, \Sigma_{\theta_\beta(s)|\text{rest}}\right)$$

$$\Sigma_{\gamma_k(s)}|\text{rest} = \left(\frac{1}{\sigma_\gamma^2}D^{-1} + \sum_{i=1}^{n}\frac{1}{\sigma_Y^2}Z_{i,k}^2 I_L\right)^{-1}$$

$$\theta_{\gamma,k}(s)|\text{rest} \sim N\left(\Sigma_{\theta_{\gamma_j}(s)|\text{rest}}\left\{\frac{1}{\sigma_Y^2}\sum_{i=1}^{n}Z_{k,i}\left\{Y_i^* - \theta_{\eta,i} - \sum_{k'\neq k}\theta_{\gamma,j}Z_{k',i} - Q^T X_i\delta(s)Q\theta_\beta\right\}\right\}, \Sigma_{\theta_{\gamma_j}(s)|\text{rest}}\right)$$

$$\delta(s)|\text{rest} = 1\times P_1(s) + 0\times P_0(s)$$

$$P_1 \propto \exp\left\{-\frac{1}{2\sigma_Y^2}\left[\delta^T\text{diag}\left\{\beta^2\|X_i\|_2^2\right\}\delta - 2\left[\sum_i^n X_i(Y_i - \eta_i - \sum_{k=1}^{m}\gamma_j Z_{i,k})\right]^T\text{diag}\left\{\beta\right\}\delta\right]\right\}$$

Denote $D_\delta := Q^T\text{diag}\left\{\delta\right\}Q$. The log-Likelihood w.r.t. θ_β can be expressed as

$$\begin{aligned}\log L(\theta_\beta) &= -\frac{1}{2\sigma_Y^2}\sum_{i=1}^{n}\|Y_i^* - \theta_{\eta,i}^*(s) - X_i D_\delta\theta_\beta\|_2^2 \\ &= \frac{1}{\sigma_Y^2}\left(\sum_{i=1}^{n}(Y_i^* - \theta_{\eta,i}(s) - \sum_{k=1}^{m}\theta_{\gamma,k}Z_{i,k})X_i\right)^T D_\delta\theta_\beta - \frac{\sum_{i=1}^{n}X_i^2}{2\sigma_Y^2}\theta_\beta^T D_\delta^2\theta_\beta \qquad (16)\end{aligned}$$

Based on the above derivations, we pre-compute the following summary statistics. These notations will help readers understand our code if one is interested.

- `XY_term_allsample`: $\sum_{i=1}^{n}X_iY_i(s) \in \mathbb{R}^p$
- `XqYstar_term_allsample`: $Y_{L\times n}^* Z_{n\times m} \in \mathbb{R}^{L\times m}$
- `XXq_sumsq`: $\sum_{i=1}^{n}X_iZ_i \in \mathbb{R}^m$
- `XcXq_sumsq`: $\left(\sum_{i=1}^{n}Z_{i,j}Z_{i,[-k]}\right)_{j=1}^{q} \in \mathbb{R}^{(m-1)\times m}$
- `XY_eta_term`: $\sum_{i=1}^{n}X_iY_i(s) - \eta_{p\times n}X_{n\times 1} \in \mathbb{R}^{p\times 1}$

- `XqYstar_theta_eta_term`: $Y^*_{L\times n} - \theta_{\eta,(L\times n)} \in \mathbb{R}^{L\times n}$

From the above derivation, we can see that the posterior covariance matrix for θ_β is a dense matrix, hence we only apply the SGLD algorithm on θ_β.

A.2 Imputing missing outcome with individual-mask

In this subsection, we consider how to update the missing outcome $Y_i(s_j)$ when considering the individual masks. Denote Q_i to be the i-th rows in Q. For subject i with missing voxel $s_j \notin \mathcal{V}_i$,

$$Y_i(s_j) = X_i\delta(s_j)Q_j\theta_\beta + \sum_{k=1}^{q} Q_j\theta_{\gamma,k}Z_{i,k} + Q_j\theta_{\eta,i} + \epsilon_i \tag{17}$$

In practice, to avoid repeated access of the entire data set, we save all locations of missing voxels in a vector `Y_imp`. When accessing one batch of outcome data, the corresponding missing locations in $Y_i(s)$ is replaced by the imputed data saved in `Y_imp`.

Below is the detail algorithm for the imputation part. We use $\mathcal{V}_i^c$ to denote the collection of all missing voxels for subject i.

Note that when computing the likelihood, we need the imputed values from `Y_imp`. The reverse mapping from individual i location s_j to a particular k-th element in `Y_imp` is similar to the above algorithm.

B Additional Simulation Results

Below, Fig 9 provides the simulation results in Simulation I with missing Pattern II. The TPR are generally greater than with missing Pattern I since the missing pixels in missing Pattern II are all located outside of the active region. MUA presents the worst performance and SBIOSimp has the best performance across all settings, but the difference between

Algorithm 2 Update missing outcomes at one iteration for subjects in the b batch.

Let `Y_imp` be the vector $\{Y_i(s_j) : s_j \in \mathcal{V}_i^c, i = 1, \ldots, n, j = 1, \ldots, p\} \in \mathbb{R}^{\sum_{i=1}^n |\mathcal{V}_i^c|}$. The length of `Y_imp` is the total number of missing outcomes across subjects and across locations.
Let $\mathcal{M}$ be a list of the index map that stores the location of missing voxels in `Y_imp`, i.e. the missing voxel $Y_i(s_j)$ is located at the k-th element in `Y_imp`.
for region $r = 1, \ldots R$, **do**
    Extract index set $\mathcal{V}_r^c$, a list of index vectors, the i-th element is the vector $\mathcal{V}_i^c \cap \mathcal{B}_r$.
    Extract index set $\mathcal{M}_r$, a list of index vectors, the i-th element is the vector $\mathcal{V}_i^c \cap \mathcal{M}$.
    for subject i with in batch b **do**
        Extract index vector $\mathcal{V}_{r,i}^c$ which is the ith element in $\mathcal{V}_r^c$
        Extract index vector $\mathcal{M}_{r,i}$ which is the ith element in $\mathcal{M}_r$
        if $\mathcal{V}_{r,i}^c$ and $\mathcal{M}_{r,i}$ are not empty set **then**
            Update `Y_imp` at locations $\mathcal{M}_{r,i}$ using (17) where $Y_i(s_j), s_j \in \mathcal{V}_{r,i}^c$.
        end if
    end for
end for

SBIOSimp and the other competing methods are much smaller compared to the result in Simulation I with missing Pattern I.

Fig 10 provides the simulation results in Simulation I with missing Pattern II when applying Morris FDR control method to the Bayesian methods.

Fig 11 provides additional results on the reading memory in Simulation II: Memory Usage. Here, both SBIOS0 and SBIOSimp read B batches of data in loops, and save them as file-backed matrices, hence the memory usage increase for reading memory is smaller compared to BIOS and MUA.

C Additional UKBiobank Results and Figures

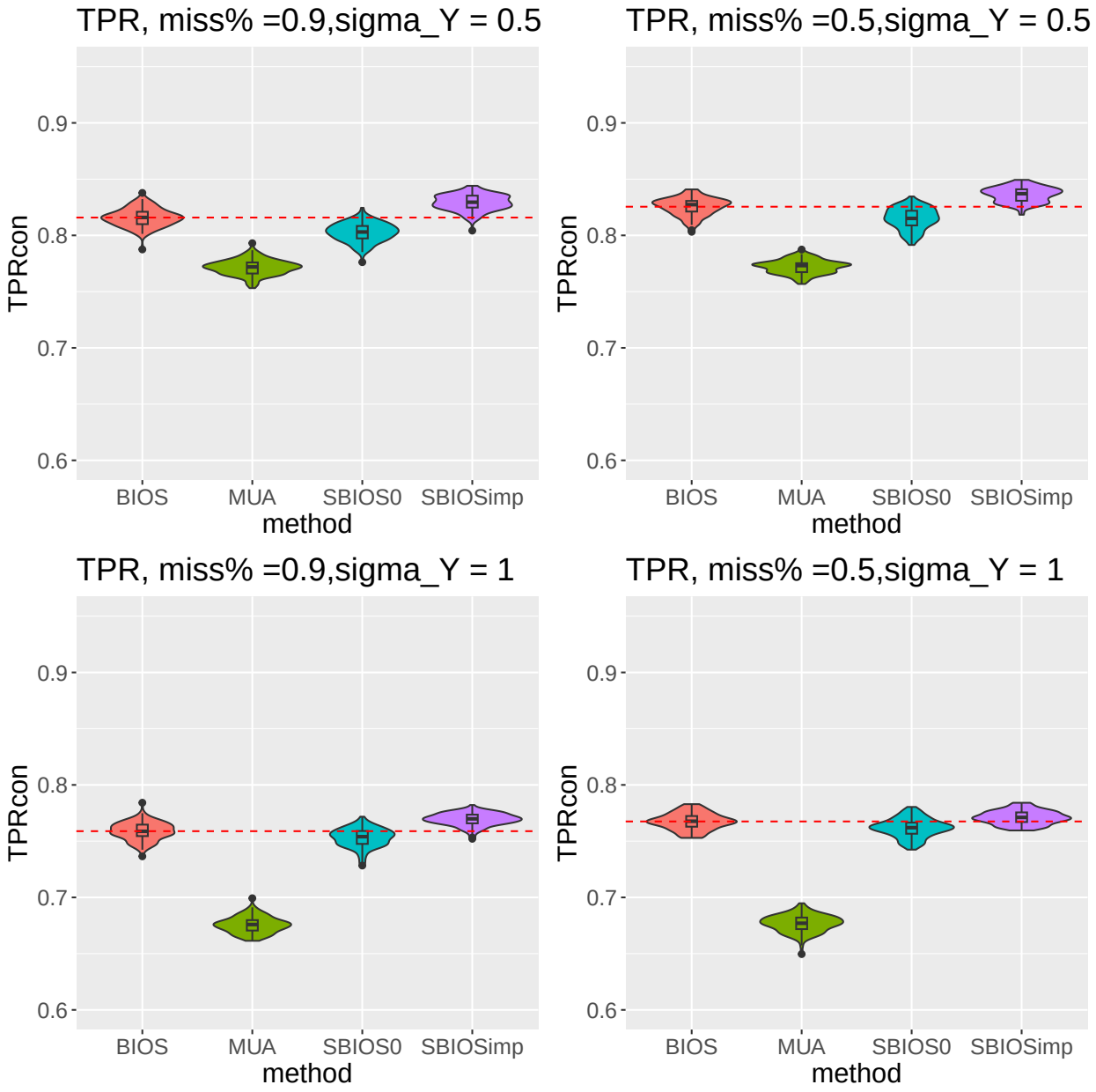


Figure 9: True Positive Rate based on 100 replicated simulation results when False Positive Rate is controlled at 0.1 when $n = 6000, p = 8100$. Missing Pattern II.

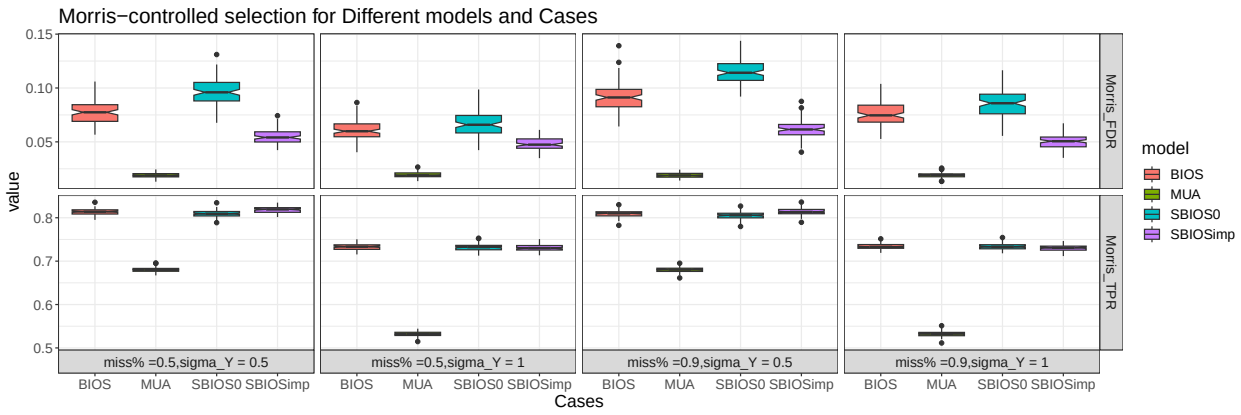


Figure 10: True Positive Rate and False Discovery Rate using Morris FDR control method, based on 100 replicated simulations, $n = 3000, p = 8100$. Missing Pattern II.

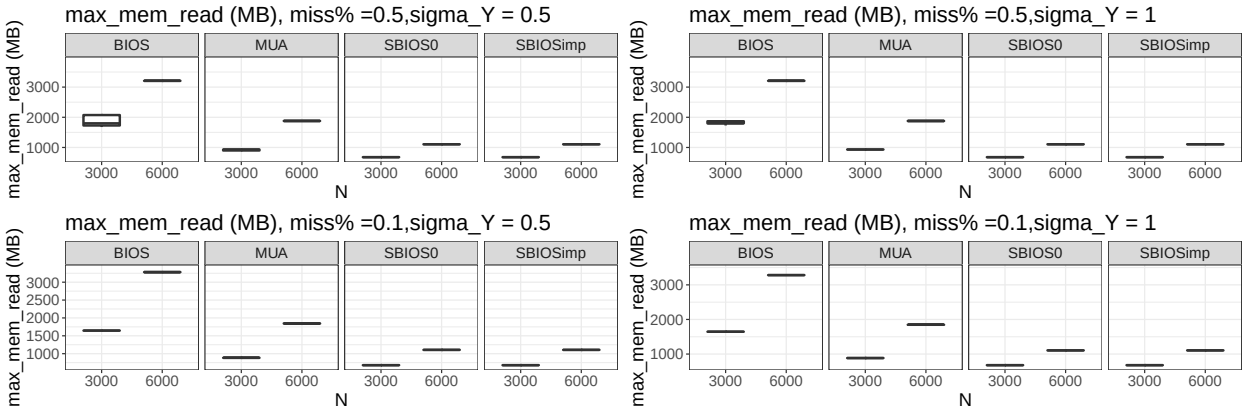


Figure 11: Maximum total reading memory (MB) for both $n = 3000$ and $n = 6000$ based on 100 replicated simulation results.

Region ID	Region Name	Case 1 and 2	Case 3	Case 4
37	Right temporal fusiform cortex, anterior division	0.97(1)	0.97(1)	0.97(1)
53	Left inferior frontal gyrus, pars triangularis	0.91(2)	0.9(2)	0.92(2)
116	right amygdala	0.89(3)	0.88(3)	0.89(3)
14	Right inferior temporal gyrus, anterior division	0.86(4)	0.87(4)	0.89(4)
115	right hippocampus	0.78(5)	0.78(5)	0.83(6)
85	Left temporal fusiform cortex, anterior division	0.75(6)	0.71(7)	0.85(5)
11	Right middle temporal gyrus, anterior division	0.73(7)	0.7(8)	0.77(7)
5	Right inferior frontal gyrus, pars triangularis	0.73(8)	0.66(10)	0.76(8)
25	Right frontal medial cortex	0.7(9)	0.72(6)	0.72(9)
62	Left inferior temporal gyrus, anterior division	0.66(10)	0.67(9)	0.72(10)

Table 3: Sensitivity Analysis Results ranked by active ratio. The active ratio for each region is reported, with the rank (decreasingly ordered by active ratio) of each region inside the brackets.

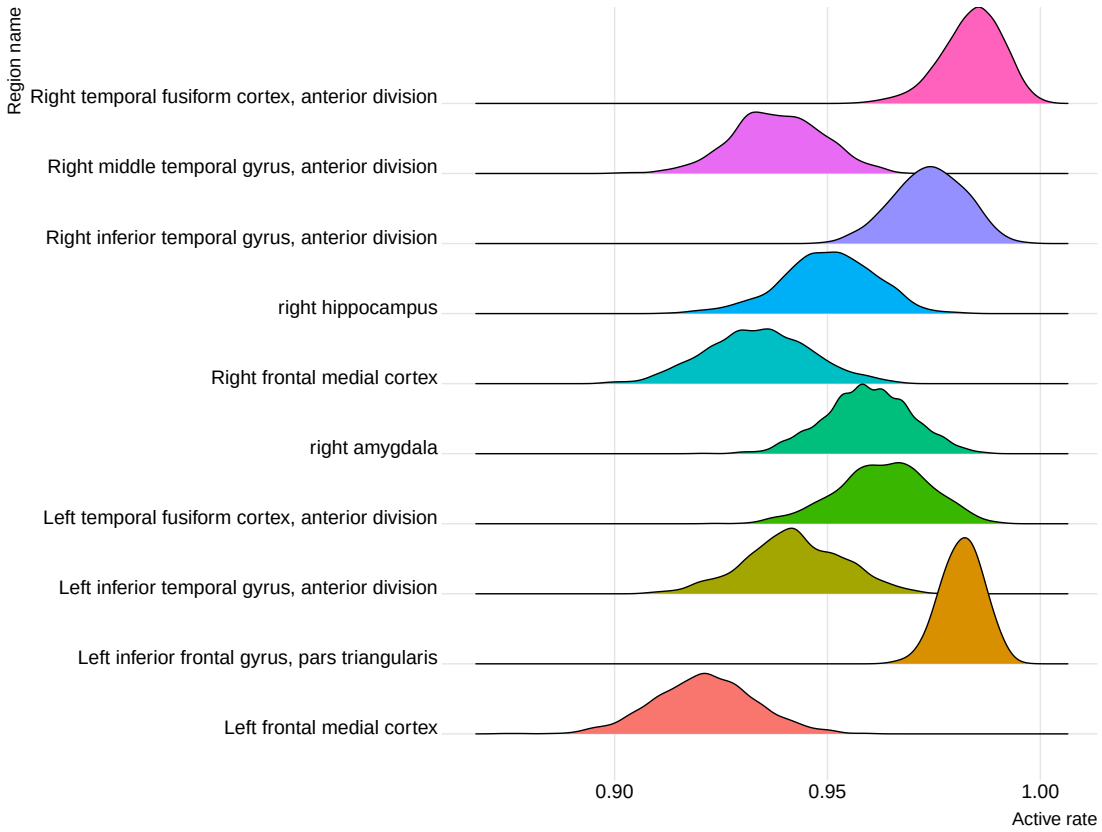
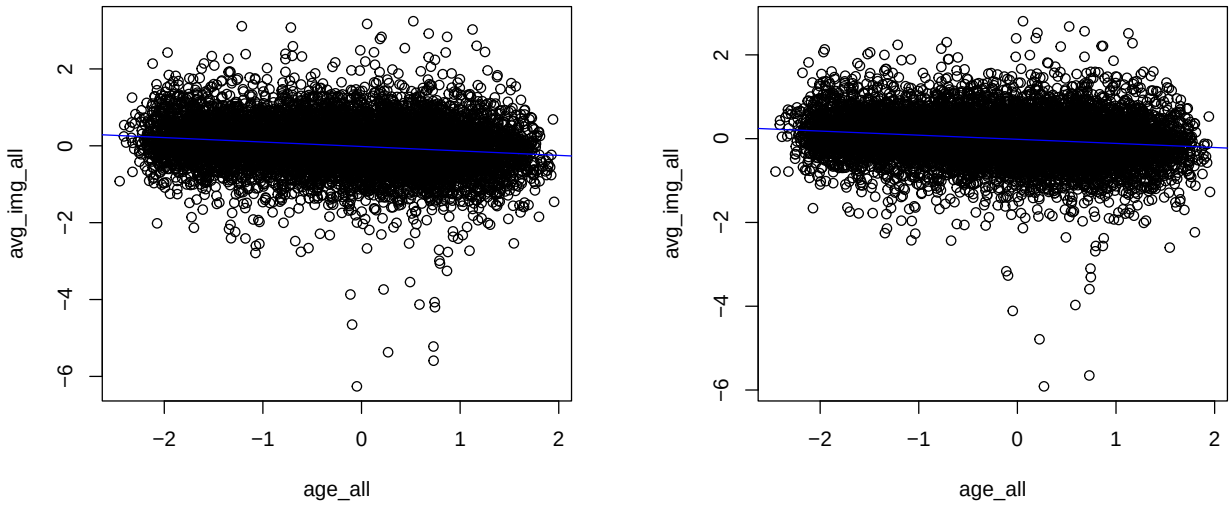


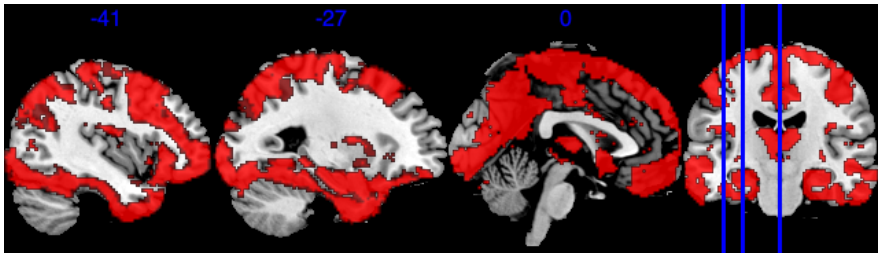
Figure 12: Posterior distribution of the region level active rate (RLAR) for the top 10 regions with highest mean region active rate.



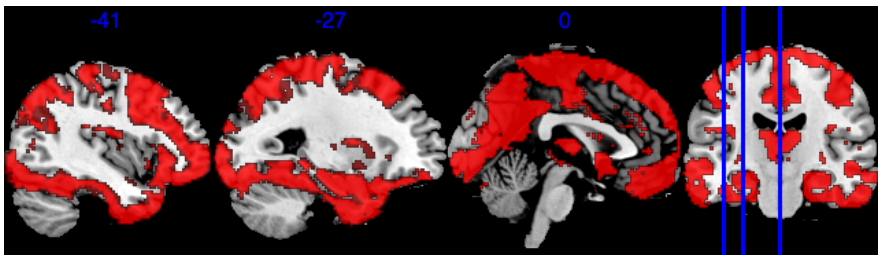
SBIOS0			
est.	Lower CI	Higher CI	p.value
-0.1731	-0.1938	-0.1523	< 0.001

SBIOSimp			
est.	Lower CI	Higher CI	p.value
-0.1666	-0.1873	-0.1457	< 0.001

Figure 13: Scatter plot of average image intensity(standardized) over the largest connected area against Age(standardized) for all individuals. The connected area is defined as follows: first we apply 0.7 cutoff on the inclusion probability to select active voxels with IP>0.7, then we search for the largest area with most connected active voxels. The scatter plot on the left is based on the inclusion probability obtained from SBIOS0 method, which happens to be a subset of the largest connected area obtained based on SBIOSimp. The right plot is based on GS-imputation.



SBIOS0 mapping: Morris FDR control with target FDR at 0.1



SBIOS mapping: Morris FDR control with target FDR at 0.1

Figure 14: Morris FDR control results for both SBIOS0 and SBIOSimp.