HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2026 May 28.

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

Yuliang Xu[a], **Timothy D. Johnson**[b], **Thomas E. Nichols**[c], **Jian Kang**[b]

[a]Department of Statistics, University of Chicago, Chicago, IL;

[b]Department of Biostatistics, University of Michigan, Ann Arbor, MI;

[c]Big Data Institute, University of Oxford, Oxford, UK

Abstract

Bayesian Image-on-Scalar Regression (ISR) provides flexible, uncertainty-aware neuroimaging analysis. However, applying ISR to large-scale datasets such as the UK Biobank is challenging due to intensive computational demands and the need to handle subject-specific brain masks rather than a common mask. We propose a novel Bayesian ISR model that scales efficiently while accommodating these inconsistent masks. Our method leverages Gaussian process priors with salience area indicators and introduces a scalable posterior computation algorithm using stochastic gradient Langevin dynamics combined with memory mapping. This approach achieves linear scaling with subsample size and constrains memory usage to the batch size, facilitating direct spatial posterior inferences on brain activation regions. Simulation studies and analysis of UK Biobank task fMRI data (38,639 subjects; over 120,000 voxels per image) demonstrate a 4- to 11-fold speed increase and an 8–18% enhancement in statistical power compared to traditional Gibbs sampling with zero-imputation. Our analysis reveals a subregion of the amygdala where emotion-related brain activation decreases by approximately 58% between ages 50 and 60. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

Keywords

Image-on-Scalar regression; Individual-specific masks; Memory-mapping; Scalable algorithm; UK Biobank data

✉**CONTACT** Jian Kang jiankang@umich.edu Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109.

Disclosure Statement

The authors have no conflicts of interests to declare.

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 (functional brain tissue) 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 article, 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 using Stochastic Gradient Langevin Dynamics (Welling and Teh 2011, SGLD) 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 used 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, Urbach, and Kutas 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 a 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 principal components idea to use spatial correlation, Ramsay and Silverman (2005) and Reiss, Huang, and Mennes (2010) proposed a penalized regression method using basis expansion to reduce the dimension of the functional outcome. Zhu, Fan, and Kong (2014) propose another model that uses local polynomial and kernel methods for estimating the spatially varying coefficients with discontinuity jumps. Yu et al. (2021) and Li et al. (2021) use bivariate spline functions to estimate the spatial functional estimator supported on the two-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 frequentist methods is that it can be difficult to make inferences on the active area selection based on these penalized low-rank models. Zhang et al. (2023) develop 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, Li, and Vannucci (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. However, the dense covariance matrix can require large memory and computational power when the outcome is very high-dimensional. To address these limitations, 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 inferences 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 where all subjects have data. However, in the UK Biobank, an intersection mask of 38,639 subjects reduces the analysis volume by 71% relative to the average subject mask volume. Some authors will attempt to minimize 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. Different from the missing data literature where the missing values exist in reality but are unobserved,

the missing values here are due to the differently aligned individual masks. The goal is to augment individual image data to a common brain mask through imputation methods. 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. A concurrent work (Lu et al. 2025) introduced an imputation approach for imaging data based on the conditional distribution of missing voxels given the observed voxels. While this represents a useful contribution, their framework assumes a common set of missing voxels across all subjects, whereas in our setting the pattern of missingness varies across individuals. In addition, their method has so far been demonstrated on datasets of more modest scale (around 1000 voxels), which differs from the much larger-scale applications we consider.

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. In particular, voxels above the nasal cavity and above the ear canals suffer from signal loss and are classified as non-brain tissue in a subject-specific manner.

All brain image analyses require 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 the most 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, Rachel Wang, and Wong (2022) proposed a Metropolis-Hasting algorithm using mini-batches where the proposal and acceptance probability are both approximated by the current mini-batch. Kim, Song, and Liang (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, Anwander, and Knösche 2008; Kulkarni, Merchant, and Awate 2022), and various scalable extensions of variational inference (Hoffman et al. 2013; Ranganath, Gerrish, and Blei 2014; Blaiotta, Cardoso, and Ashburner 2016). Although these variational methods are computationally efficient, 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.

This article is organized as follows. Section 2 introduces our Bayesian Image-on-Scalar model; Section 3 details the algorithm and computational aspects; Section 4 applies the method to the UK Biobank imaging data and presents sensitivity analyses; and Section 5 summarizes our contributions and discusses our findings. We defer the simulation results to supplementary Section S2. Additional real data analysis results are provided in the supplementary materials. Our implementation is provided as an R package, SBIOS, and is publicly available on GitHub.[1]

2. Model

Let $\mathscr{B} \subset \mathbb{R}^3$ denote the entire brain region. Let $\{s_j\}_{j=1}^p \subset \mathscr{B}$ be a set of fixed grid points in $\mathscr{B}$, 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 $\mathscr{V}_i$ denote the set of locations where the image intensity is observed for individual i, that is, for any $s_j \in \mathscr{V}_i, Y_i(s_j)$ is not missing. For any $i = 1, \ldots, n, \mathscr{V}_i \subset \{s_1, \ldots, s_p\}$. Let X_i be the primary covariate of interest, Z_{ik} be the kth confounding covariate for $k = 1, \ldots, q$. We propose an Image-on-Scalar regression model. For individual i and any $s_j \in \mathscr{V}_i$,

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

[1] See the SBIOS R package on Github page https://github.com/yuliangxu/SBIOS or in the supplementary

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 kth 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, Fan, and Kong (2014).

For model (1), we specify the following priors:

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

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

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

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

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. The priors in (2)–(5) are mutually independent on the prior level. Popular choices of the kernel function κ include the exponential square kernel and the Matérn kernel (6).

$$\kappa(s', s; \upsilon, \rho) = C_\upsilon\left(\|s' - s\|_2^2 / \rho\right),$$

$$C_\upsilon(d) \coloneqq \frac{2^{1-\upsilon}}{\Gamma(\upsilon)} (\sqrt{2\upsilon} d)^\upsilon \, K_\upsilon\left(\sqrt{2\upsilon} d\right), \tag{6}$$

where K_ν is a modified Bessel function of the second kind (Rasmussen and Williams 2005). We use the Matérn kernel 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 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 4.

Model (1) can be extended to include a selection variable δ for multiple terms.

$$Y_i(s_j) = \sum_{h=1}^{H} X_{i,}\beta_h(s_j)\delta_h(s_j) + \sum_{k=1}^{m} \gamma_k(s_j)Z_{ik} + \eta_i(s_j) + \epsilon_i(s_j), \quad \epsilon_i(s_j) \sim \mathrm{N}(0, \sigma_Y^2). \tag{7}$$

In Section 4.4.2, we provide an analysis on the UKB data where a common $\delta(s)$ is applied to the main effect and the interaction effect. For the rest of this article, we follow model (1) and only view model (7) as an extended framework.

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,

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

where λ_l is the lth eigenvalue, and ψ_l is the lth eigenfunction (see sec. 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 $L^2(\mathcal{B})$, that is, $\int_{s \in \mathcal{B}} \psi_l(s)\psi_{l'}(s)ds = I(l = l')$, 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 $\gamma_k(s) = \sum_{l=1}^{\infty} \theta_{\gamma,k,l}\psi_l(s)$ and $\eta_i(s) = \sum_{l=1}^{\infty} \theta_{\eta,i,l}\psi_l(s)$, 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 $\sum_{l=1}^{L}\theta_{\beta,l}\psi_l(s)$. 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 $\mathscr{B}$, denoted as $\mathscr{B}_1, \ldots, \mathscr{B}_R$. For the three GP priors (3)–(5), we assume that the kernel function $\kappa(s_j, s_k) = 0$ for any $s_j \in \mathscr{B}_r, s_k \in \mathscr{B}_{r'}, r \neq r'$, and the prior covariance matrix on the fixed grid has a block diagonal structure.

For the rth region, let p_r be the number of voxels in $\mathscr{B}_r$ and let $Q_r = \left(\psi_l(s_{r,j})\right)_{l=1,j=1}^{L_r,p_r} \in \mathbb{R}^{p_r \times L_r}$ be the matrix with the (j,l)th component $\psi_l(s_{r,j})$ where $\{s_{r,j}\}_{j=1}^{p_r}$ forms the fixed grid in $\mathscr{B}_r$. With the region partition, the GP priors on the rth region can be reexpressed as $\boldsymbol{\beta}_r = \left(\beta(s_{r,1}), \ldots, \beta(s_{r,p_r})\right)^T \approx Q_r\boldsymbol{\theta}_{\beta,r}$, where $\boldsymbol{\theta}_{\beta,r} \sim \mathcal{N}(0, \sigma_\beta^2 D_r)$ and D_r is a diagonal matrix with diagonal $\left(\lambda_{r,1}, \ldots, \lambda_{r,L_r}\right) \in \mathbb{R}^{L_r}$. Note that Q_r is not necessarily orthonormal as the finite approximation of the eigenfunctions, hence, in practice we apply QR decomposition on the matrix formed by eigenfunctions, and take the Q-matrix as the final approximation of Q_r to guarantee orthonormality.

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

$$\mathbf{Y}_{i,r}^* = Q_r^T X_i \text{diag}\{\boldsymbol{\delta}_r\} Q_r\boldsymbol{\theta}_{\beta,r} + \sum_{k=1}^{m}\theta_{\gamma,k,r}Z_{i,k} + \theta_{\eta,i,r} + \boldsymbol{\epsilon}_{i,r}^* \tag{8}$$

with the prior specification $\boldsymbol{\theta}_{\beta,r} \sim N(0, \sigma_\beta^2 D_r), \boldsymbol{\theta}_{\gamma,k,r} \sim N(0, \sigma_\gamma^2 D_r), \boldsymbol{\theta}_{\eta,i,r} \sim N(0, \sigma_\eta^2 D_r)$, and the low-dimensional noise $\boldsymbol{\epsilon}_{i,r}^* \sim N(0, \sigma_Y^2 I_{L,r})$. The working model (8) 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 (6) and sensitivity analysis on the 90% cutoff are discussed in Sections 4 and 4.4.1.

3.2. Scalable Algorithm for a Large Dataset

Inspired by Welling and Teh (2011) and Wu, Rachel Wang, and Wong (2022), we propose Algorithm 1 based on SGLD to sample the posteriors from a large dataset. The SGLD algorithm outperforms Gibbs sampling (GS) in three ways: (a) under the assumption that all individuals are independently 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; (b) 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 (c) 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 tth iteration. Let τ_t be the step size at the tth iteration, let $\pi(\theta_{\beta,l})$ denote the prior, and let $\pi_{i\in\mathcal{J}}(Y_i \mid X_i, Z_i, \theta) := \prod_{i\in\mathcal{J}}\pi(Y_i \mid X_i, Z_i, \theta)$ be the likelihood for a subsample $\mathcal{J}$, where θ is the collection of all parameters. Denote by n_s the number of subjects in the subsample $\mathcal{J}$. Let $\nabla f(x)$ be the gradient of $f(x)$. At the tth iteration, the lth component in θ_β is updated as

$$\theta_{\beta,l}^{(t)} \longleftarrow \theta_{\beta,l}^{(t-1)} + \frac{\tau_t}{2}\nabla \log \pi\left(\theta_{\beta,l}^{(t-1)}\right) + \frac{\tau_t}{2}\frac{n}{n_s}\nabla \log \pi_{i\in\mathcal{J}}\left(Y_i \mid X_i, Z_i, \theta^{(t-1)}\right) + \sqrt{\tau_t}\varepsilon_l \tag{9}$$

where $\varepsilon_l \overset{\text{iid}}{\sim} N(0,1)$. Let L_r be the number of basis coefficients for the rth brain region. The time complexity for (9) 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 1, where the large dataset 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.

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 tth iteration.

5: Load batch b into memory.

6: **for** region $r = 1, \ldots, R$ **do**

7: Randomly select a subsample $\mathcal{I}$ of size n_s from batch b.

8: Update $\theta_{\beta,l}$ for region r using (9) 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**

On line 10 of Algorithm 1, using Gibbs sampling to update γ, δ also requires the entire 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 dataset. 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 (8). 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 1, 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 $\eta_i(s)$ is more of a nuisance parameter and not our main 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.

For the optional imputation found on line 15 of Algorithm 1, 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 $Y_i(s_j), s_j \in \mathscr{V}_i^c$ is updated. The time complexity to update all missing values in Y_i is $O(L \times (|\mathscr{V}_i^c|))$ where $|\mathscr{V}_i^c|$ 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 supplementary materials.

Algorithm 1 is implemented using the Rcpp package `bigmemory` (Kane, Emerson, and Weston 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. At the beginning of Algorithm 1, 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 is passed to Algorithm 1, accessing different batches of data becomes very efficient and memory-conserving.

3.3. 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).[2]

3.4. Ablation Study Design

To demonstrate the performance of the proposed imputation and computation schemes, in supplementary Section S2, we compare our proposed model in three variations, BIOS, SBIOS0, and SBIOSimp, with the existing method MUA in three simulation examples. Note that BIOS, SBIOS0, and SBIOSimp are based on the same model (1) with the same set of priors. The difference is that, BIOS uses a fully Gibbs-sampler without the advanced computational techniques (SGLD, memory-mapping, etc.) with 0 imputation; SBIOS0 uses advanced computation techniques with 0 imputation; and SBIOSimp uses advanced techniques with PCA-based imputation (which we refer to as Gibbs-sampler imputation approach). From BIOS, SBIOS0, to SBIOSimp, we added one advanced feature at a time

[2]Code for all simulations and all implementations of the proposed methods and competing methods can be found in the R package SBIOS (https://github.com/yuliangxu/SBIOS).

as shown in Figure 1. We implemented and compared all three methods with the baseline method MUA in Simulations I–III in supplementary Section S2, serving as an ablation study to systematically remove or modify components of an algorithm and understand the contribution of each component.

The three simulation results in supplementary Section S2 show the superior performance of SBIOSimp in three aspects: selection accuracy (Simulation I), maximum memory usage (Simulation II), and time scalability (Simulation III).

4. UK Biobank Application

In this real data application, we use 3D task fMRI data from the UK Biobank as the outcome and age as the single exposure variable. The three confounding variables are gender, age by gender interaction, and head size. The original range of age is from 44 to 83. We use the standardized age (standardized by $\{X_i - \bar{X}\}/\mathrm{SD}(X)$ where $\bar{X}$ and $\mathrm{SD}(X)$ stands for the sample mean and standard deviation respectively) as the exposure. Gender is coded as a binary variable with 0 being female and 1 being male. The interaction of age by gender is computed using the standardized age times gender. The parameter for the interaction term, $\gamma_{\mathrm{age}\times\mathrm{gender}}(s)$, represents the standardized age effect for males minus the age effect for females.

4.1. Data Preprocessing and Estimation Procedure

The outcome variable is the fMRI data obtained from an emotion recognition task. In this task, participants are required to identify which of two faces displaying fearful expressions (or shapes) presented at the bottom of the screen matches the face (or shape) displayed at the top of the screen (See details in the Hariri faces/shapes emotion task (Hariri et al. 2002; Miller et al. 2016), as implemented in the Human Connectome Project). The 3D fMRI data in MNI atlas space (Evans et al. 1994) is a rectangular prism of 91×109×91 voxels, and we used a total of $n = 38{,}639$ subjects with task fMRI. The NIFTI data is approximately 180 GB, and the processed RDS files are roughly 34 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 the R package `RNifti` (Clayden, Cox, and Jenkinson 2023), the preprocessing time for one subject's data, including directly loading the outcome and mask data from a DropBox folder, is around 2.3 sec. Saving the preprocessed data into file-backed matrices is instantaneous. We use the Harvard-Oxford cortical and subcortical structural atlases (Desikan et al. 2006) for the analysis. After preprocessing, we have a total of 110 brain regions.

To define the analysis mask, we recall the definition of observed proportion (OP) at location s_j to be $h(s_j) = n^{-1}\sum_{i=1}^{n} 1_{\mathscr{V}_i}(s_j)$, where $\mathscr{V}_i$ is the set of all observed locations for individual i. We define the group analysis mask as $\mathscr{B} = \{s_j : h(s_j) > 0.5\}$, that is, the area where each voxel has at least 50% observed data. As shown in Figure 2, the group analysis mask with completely observed data (purple area) covers significantly less area compared to $\mathscr{B}$, which has at least 0.5 observed proportion (blue area). In the complete observed data, large portions of the brain regions are missing, notably including the orbitofrontal cortex,

the inferior temporal cortex, and the amygdala—regions crucial for emotion processing. In particular, the mask with complete observed data contains only 52 out of the 110 regions in the Harvard-Oxford atlas.

After applying a common mask $\mathscr{B}$ with an observed proportion of 0.5, we end up with $\mathscr{B}$ that contains $p = 121{,}865$ voxels. The image outcomes $Y_i(s)$ are standardized across subjects, that is, $Y_i(s) = \{M_i(s) - \bar{M}(s)\}/\mathrm{SD}[\mathrm{M(s)}]$ where $M_i(s)$ is the original image for subject i location s, $\bar{M}(s)$ is the sample mean, and SD[M(s)] is the standard deviation of $\{M_i(s)\}_{i=1}^{n}$. For each region, we apply the Matérn kernel function but with different ρ and υ parameters (6), to account for the different smoothness of each region. Both ρ and υ are determined through grid search so that the empirical covariance of $Y(s_j), Y(s_j^{'})$ and the estimated covariance by the Matérn kernel have the smallest difference in Frobenius norm. The number of bases L is chosen so 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 $L = \sum_{r=1}^{110} L_r = 16{,}879$. In Section 4.4.1, we provide a sensitivity analysis when the cutoff is based on 92% of the total summation.

The $n = 38{,}639$ subjects are split into 50 batches of data, with each batch containing around 700–800 subjects. The subsample size is 200, and the step size decay parameters $a = 0.0001, b = 1, \gamma = 0.35$ are chosen so that the step size roughly decreases from 7×10^{-5} to 5×10^{-6} over the 5000 MCMC iterations. We use the MUA results as the initial values for β and γ to run the SBIOS0 and SBIOSimp methods. The initial values for $\theta_{\eta,i,l}$ are set to 0 everywhere, and the initial values for $\delta(s)$ are set to 1 everywhere. The initial values for $\sigma_Y, \sigma_\eta, \sigma_\beta, \sigma_\gamma$ are all set to 0.1. To check the convergence of SBIOSimp, we run three independent chains and use the Gelman and Rubin test. The l_2 norm of the residuals is computed for the last 1000 iterations from the three chains using 20% of the entire data. The point estimate of the Gelman-Rubin statistic is 1.03 with an upper confidence limit of 1.1, indicating that the MCMC chain has approximately converged.

4.2. Analysis Results

4.2.1. Age-Related Emotion Recognition Brain Activation Patterns: We present the top 10 regions identified by the SBIOSimp method in Table 1. This SBIOSimp result takes 20.5 hr to run on 1 core CPU HPC Cluster, and the max memory used is 14 GiB. Using the posterior sample for each region, we compute the Region Level Activation Rate (RLAR) as follows: denote $\mathscr{B}_j$ as the mask for region j, the RLAR is $\sum_{s \in \mathscr{B}_j} \delta(s)/|\mathscr{B}_j|$, where $|\mathscr{B}_j|$ is the total number of voxels in region j. Hence, for each MCMC sample of $\delta(s)$, we 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 Section S4 in the supplementary material. 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 effects. Only voxels with a marginal posterior inclusion probability (PIP) over 95% are viewed as active

voxels. Because the top 10 regions have 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 a negative effect, that is, as age increases, the brain signal intensity on selected voxels tends to decrease. This trend is also reflected in the raw data, as can be seen by the scatter plots of age against the average image intensity in Section S4 in the supplementary material. To interpret the numeric results in Table 1, take the *Right temporal fusiform cortex, anterior division* region as an example. On average, 99% of the voxels within this region have activity that is associated with age. Specifically, there is a decrease in brain signal intensity of 14.81 standard deviations summed over all locations within this region for 1 standard deviation increase in age. The last column represents the median percentage decline in the brain signal intensity if age increases from 50 to 60. For *Right temporal fusiform cortex, anterior division*, this means when age increases from 50 to 60, the median decline in the brain signal intensity among all voxels in this region will be 75.87%. Supplementary Section S4.1 provides the mathematical definition and detailed derivation for this percentage.

For a visual representation, Figure 3(a) shows the voxel level PIP in the sagittal plane. The highlighted red regions represent voxels with greater than 0.95 PIP. Figure 3(b) presents the effect size of $\beta(s)\delta(s)$, with the highlighted area in the range $(-0.06, -0.03)$. Note that from Figure 3(b), voxels with PIP greater than 0.95 also correspond to voxels with a larger absolute value of effect size. We notice that the activation region (defined by voxel level PIP greater than 0.95) has a negative effect $\beta(s)$. This can also be validated by the scatterplot in Section S4 in the supplementary material, where the image intensity generally has a negative association with age across all individuals.

Based on our results, we have the following general interpretations: (i) when controlling 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 3(b), indicating a very low voxel level signal-to-noise ratio; (iii) the top five brain regions with the highest RLAR are (a) *right intracalcarine cortex, right supracalcarine cortex*, and *left Temporal fusiform cortex, anterior division*, all considered as critical areas for high-level visual processing including face recognition; (b) *left temporal fusiform cortex, anterior division*, a key structure for face perception, object recognition, and language processing (Weiner and Zilles 2016); 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 five regions are also consistently identified in the sensitivity analysis when using half of the data as training data, see Section 4.4.1.

To further demonstrate the voxel-level results provided by SBIOSimp, in the next section, we perform a detailed analysis of the amygdala region.

4.2.2. Focused Results: Amygdala Region: The amygdala region is part of the limbic system. It is responsible for detecting danger and negative emotions, and play an important role in behavior, emotional control, and learning (Bzdok et al. 2013). The emotion task fMRI data in UK Biobank is based on emotion tasks where participants are asked to identify faces with negative emotions. Hence, we expect the amygdala region to play an

important role. The amygdala region is a small area in the brain, as shown in Figure 4(a), containing 380 voxels out of a total of 121,865 voxels.

Based on the results shown in Figure 4, SBIOSimp identifies a large proportion of voxels in the amygdala region to be active. Numerically, SBIOSimp identifies 324 out of 380 voxels in both the left and right amygdala to be active with PIP > 0.95. From the highlighted box in Figure 4(b), the active voxels mostly concentrate on the direction where the amygdala region connects to the *parahippocampal gyrus, anterior division*. The anterior portion of the parahippocampal gyrus is involved in complex emotive processes and has significant interconnectivity with other cortical limbic structures and the amygdala (Kaas 2016). Figure 4(c) shows the yellow-shaded area within the amygdala in which voxels are associated with at least 50% decline in the brain signal intensity for 10-year increase in age from 50.

4.3. Comparing SBIOSimp with SBIOS0 and MUA

We compare the posterior mean of $\beta(s)$ given $\text{PIP}(s) \geq 0.95$ between SBIOS0 and SBIOSimp stratified by the observed proportions on these regions in Figure 5. Each point in Figure 5 represents the effect of age on the brain signal for one voxel in the brain. The 6 regions in Figure 5 are chosen with high missingness. Comparing blue dots (low observed proportion, $h(s_j) \in [0.5,0.7)$), red dots (medium observed proportion, $h(s_j) \in [0.7,0.9)$) with black dots (high observed proportion, $h(s_j) \in [0.9,1]$), we can see that β fitted with SBIOS0 on voxels with lower observed proportion 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 S2.

A comparison of active β selection between MUA and SBIOSimp is available in Table 2. For SBIOSimp, the active voxel selection is based on PIP greater than 95%. For MUA, for a fair comparison, the cutoff on BH-adjusted p-values is determined using the same proportion of active voxels selected by SBIOSimp, so that MUA and SBIOSimp select the same number of active voxels. Table 2 shows that MUA with 0-imputation tends to map more voxels of low observed proportion toward 0.

4.4. Sensitivity Analysis

4.4.1. Hyperparameter Specifications: To further validate the result reported in Section 4.2.1, 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 bases in the Gaussian kernel. The baseline setting for results in Section 4.2.1 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 ($L = 16{,}879$). 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 bases based on 92% of total eigenvalues ($L = 20{,}355$). For case 5, we vary the smoothness parameter υ in the Matérn kernel (6), and create a new kernel where ν in each region is set to 90% the value of ν in the standard kernel used in Section 4.2.1. Because Case 1 & 2 have very similar results, we report them together.

Table 3 presents the region-level difference between each sensitivity case and the final result presented in Section 4.2.1. Cases 1 and 2 produce almost the same result and are reported together. We can see that the top four regions are consistently selected with the highest RLAR.

4.4.2. Model Assumptions on Selection of Interactions: In the analysis of Section 4.2.1, we apply the selection variable $\delta(s)$ exclusively to the main effect of age. In this section, we extend the model (1) by applying $\delta(s)$ to both the main effect of age and the interaction effect of age and gender. Consequently, the interaction term is automatically set to zero whenever the corresponding main effect is not selected. The extended model can be expressed as below:

$$Y_i(s_j) = X_i\beta(s_j)\delta(s_j) + \widetilde{X}_i\widetilde{\beta}(s_j)\delta(s_j) + \sum_{k=1}^{m} \gamma_k(s_j)Z_{ik} + \eta_i(s_j) + \epsilon_i(s_j), \tag{10}$$

where $\epsilon_i(s_j) \sim N(0, \sigma_Y^2)$ and X_i is the standardized age for individual i, and $\widetilde{X}_i$ is the interaction term between age and gender. In this way, $\delta(s)$ can control the active signals in the main effect and interaction effect at the same time. Here, β and $\widetilde{\beta}$ are assigned GP priors independently and both are updated using the SGLD algorithm.

The last column in Table 4 represents the average negative effect among the active voxels, which is just the {Neg Sum}/{Count} in the previous column. In Table 4, there are four regions, *Right intracalcarine cortex, Left inferior frontal gyrus, pars triangularis, Left intracalcarine cortex*, and *Left occipital pole* have made less active voxel selection after applying a common δ to the interaction term, whereas the remaining six regions have little change.

In addition, Figure 6 shows results equivalent to those in Figure 3, but with the selection indicator $\delta(s)$ applied simultaneously to both the main effect and the interaction effect. As shown in Figure 6(b), the posterior of $\delta(s)$ in model (10) is driven by both $\beta(s)$ and $\widetilde{\beta}(s)$. The regions exhibiting strong negative effects of $\beta(s)$ do not fully align with the regions showing PIP > 0.95, in contrast to the original results in Figure 3(b). Therefore, if the primary interest is in the main exposure variable X_i, such as age in our analysis, we recommend applying $\delta(s)$ exclusively to this primary effect to achieve a more accurate selection of activation regions.

Since the gender variable is binary with female being 0 and male being 1, the interpretation for $\widetilde{\beta}(s)$ is that comparing to the female subjects, one standard deviation (s.d.) increase in age for male subjects is associated with $\widetilde{\beta}(s)$-s.d. of change in the image intensity. The boxed green area in Figure 6(c) and (b) identifies one active area where $\beta(s)$ has a negligible effect, but $\widetilde{\beta}(s)$ has a large effect size, indicating that this area is associated with the differences of male's age-brain intensity association compared to female. For example, one s.d. increase in male's age is associated with at least 0.01 s.d. increase in brain signal intensity compared to the female baseline in this green-boxed area. On the other hand,

Figure 6(d) also identifies areas where one s.d. increase in male's age is associated with at least 0.01 s.d. decrease in brain signal intensity compared to the female baseline. The area in the green box spans several brain regions in the right hemisphere, including *Right lateral occipital cortex, superior division, Right insular cortex, Right middle temporal gyrus, posterior division*, and *Right frontal operculum cortex*. They jointly integrate information from multiple modalities and detect behaviorally relevant stimuli. The negative $\widetilde{\beta}(s)$ in Figure 6(d) spans over *Right parahippocampal gyrus, posterior division, Right temporal fusiform cortex, posterior division, Left temporal fusiform cortex, posterior division, Left temporal pole*. They jointly process and integrate vision and semantic information and are related to contextual and memory functions.

5. Discussion

We propose a Bayesian hierarchical model for Image-on-Scalar regression, with a computationally efficient algorithm for posterior sampling, and apply our proposed method to the UK Biobank data. Our proposed model can capture high dimensional spatial correlation, account for the individual-level correlated noise, and provide uncertainty quantification on the active area selection through posterior inclusion probability. Our main computational contribution includes (i) employing a scalable Stochastic Gradient Langevin Dynamics (SGLD) algorithm for big data ISR, (ii) borrowing memory-mapping techniques to overcome memory limitations for large-scale imaging data, and (iii) using individual-specific masks with an imputation approach to maximize the use of available imaging data.

Extensive simulations compare our model's implementations — standard Gibbs sampling (BIOS), SGLD with missing outcomes imputed as 0 (SBIOS0), and our imputation-based SGLD (SBIOSimp) — against Mass Univariate Analysis (MUA). Results show that SBIOSimp achieves superior variable selection accuracy, lower memory cost, and scalable computation. UK Biobank application results are validated by sensitivity analysis under different hyperparameters and kernel settings. Note that although we only compare SBIOSimp with the baseline 0-imputation method since 0-imputation is the only off-the-shelf alternative imputation method to be applied on such a large scale, there exist other directions of imputation approaches, such as the functional PCA (Happ and Greven 2018).

Simulations and UK Biobank data demonstrate our method's potential for large-scale imaging, yet several limitations persist. First, our Gaussian process depends on a user-defined kernel function, and our adaptive kernel parameter selection is partially subjective. Second, for computational efficiency we assume the individual effect η_i shares the same kernel as β and γ, which may overlook individual-level latent confounders. Lastly, while SGLD accurately estimates the first moment of β, its diminishing step size may fail to capture the true variance of β.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the Editor Professor Michael Stein, the Associate Editor and reviewers for their helpful comments and constructive suggestions, which led to this much-improved manuscript.

Funding

This work was supported by National Institutes of Health grant R01DA048993. Kang's research was partially supported by National Institutes of Health grant R01MH105561 and National Science Foundation grant IIS2123777.

Data Availability Statement

The real data, UK Biobank data (Miller et al. 2016), used in Section 4, is publicly available online at https://www.ukbiobank.ac.uk/. The preprocessed data supporting the findings of this study are available from the corresponding author on request. Most of this work was completed during Yuliang Xu's PhD training in the Department of Biostatistics at the University of Michigan.

References

Blaiotta C, Cardoso MJ, and Ashburner J (2016), "Variational Inference for Medical Image Segmentation," Computer Vision and Image Understanding, 151, 14–28.

Bzdok D, Laird AR, Zilles K, Fox PT, and Eickhoff SB (2013), "An Investigation of the Structural, Connectional, and Functional Subspecialization in the Human Amygdala," Human Brain Mapping, 34, 3247–3266. [PubMed: 22806915]

Chen G, Saad ZS, Britton JC, Pine DS, and Cox RW (2013), "Linear Mixed-Effects Modeling Approach to FMRI Group Analysis," Neuroimage, 73, 176–190. [PubMed: 23376789]

Chen G, Saad ZS, Nath AR, Beauchamp MS, and Cox RW (2012), "FMRI Group Analysis Combining Effect Estimates and their Variances," Neuroimage, 60, 747–765. [PubMed: 22245637]

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

Cole JH (2020), "Multimodality Neuroimaging Brain-Age in UK Biobank: Relationship to Biomedical, Lifestyle, and Cognitive Factors," Neurobiology of Aging, 92, 34–42. [PubMed: 32380363]

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, and Killiany RJ (2006), "An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans into Gyral based Regions of Interest," Neuroimage, 31, 968–980. [PubMed: 16530430]

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 LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, Marchini J, and Smith SM (2018), "Genome-Wide Association Studies of Brain Imaging Phenotypes in UK Biobank," Nature, 562, 210–216. [PubMed: 30305740]

Evans AC, Kamber M, Collins D, and MacDonald D (1994), "An MRI-based Probabilistic Atlas of Neuroanatomy," in Magnetic Resonance Scanning and Epilepsy, pp. 263–274, New York: Springer.

Groppe DM, Urbach TP, and Kutas M (2011), "Mass Univariate Analysis of Event-Related Brain Potentials/Fields I: A Critical Tutorial Review," Psychophysiology, 48, 1711–1725. [PubMed: 21895683]

Hammers A, Asselin M-C, Hinz R, Kitchen I, Brooks DJ, Duncan JS, and Koepp MJ (2007), "Upregulation of Opioid Receptor Binding Following Spontaneous Epileptic Seizures," Brain, 130, 1009–1016. [PubMed: 17301080]

Happ C, and Greven S (2018), "Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains," Journal of the American Statistical Association, 113, 649–659.

Hariri AR, Tessitore A, Mattay VS, Fera F, and Weinberger DR (2002), "The Amygdala Response to Emotional Stimuli: A Comparison of Faces and Scenes," Neuroimage, 17, 317–323. [PubMed: 12482086]

Hoffman MD, Blei DM, Wang C, and Paisley J (2013), "Stochastic Variational Inference," Journal of Machine Learning Research 14:1303–1347.

Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, and Evans AC (1998), "Enhancement of MR Images Using Registration for Signal Averaging," Journal of Computer Assisted Tomography, 22, 324–333. [PubMed: 9530404]

Jaakkola TS, and Jordan MI (1999), "Variational Probabilistic Inference and the QMR-DT Network," Journal of Artificial Intelligence Research, 10, 291–322.

Kaas JH (2016), Evolution of Nervous Systems, Amsterdam: Academic Press.

Kaden E, Anwander A, and Knösche TR (2008), "Variational Inference of the Fiber Orientation Density Using Diffusion MR Imaging," Neuroimage, 42, 1366–1380. [PubMed: 18603006]

Kane MJ, 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. [PubMed: 35559269]

Kulkarni PH, Merchant S, and Awate SP (2022), "Mixed-Dictionary Models and Variational Inference in Task fMRI for Shorter Scans and Better Image Quality," Medical Image Analysis, 78, 102392. [PubMed: 35235896]

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. [PubMed: 34293465]

Li X, Wang L, Wang HJ, and Initiative ADN (2021), "Sparse Learning and Structure Identification for Ultrahigh-Dimensional Image-on-Scalar Regression," Journal of the American Statistical Association, 116, 1994–2008.

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

Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Oesingmann N, Alfaro-Almagro F, Bell JD, Boultwood C, Collins R, Conroy MC, et al. (2020), "The UK Biobank Imaging Enhancement of 100,000 Participants: Rationale, Data Collection, Management and Future Directions," Nature Communications, 11, 2624.

Lu T, Kochunov P, Chen C, Huang H-H, Hong LE, and Chen S (2025), "A New Multiple Imputation Method for High-Dimensional Neuroimaging Data," Human Brain Mapping, 46, e70161. [PubMed: 40116075]

Maullin-Sapey T, and Nichols TE (2022), "BLMM: Parallelised Computing for Big Linear Mixed Models," NeuroImage, 264, 119729. [PubMed: 36336314]

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JL, et al. (2016), "Multimodal Population Brain Imaging in the UK Biobank Prospective Epidemiological Study," Nature Neuroscience, 19, 1523–1536. [PubMed: 27643430]

Morris JS (2015), "Functional Regression," Annual Review of Statistics and Its Application, 2, 321–359.

Mulugeta G, Eckert MA, Vaden KI, Johnson TD, and Lawson AB (2017), "Methods for the Analysis of Missing Data in FMRI Studies," Journal of Biometrics & Biostatistics, 8, 335. [PubMed: 31080693]

Onitsuka T, Shenton ME, Salisbury DF, Dickey CC, Kasai K, Toner SK, Frumin M, Kikinis R, Jolesz FA, and McCarley RW (2004), "Middle and Inferior Temporal Gyrus Gray Matter Volume

Abnormalities in Chronic Schizophrenia: An MRI Study," American Journal of Psychiatry, 161, 1603–1611. [PubMed: 15337650]

Ramsay JO, and Silverman BW (2005), Fitting Differential Equations to Functional Data: Principal Differential Analysis, New York: 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, Vol. 33, proceedings of Machine Learning Research, eds. Kaski S and Corander J, pp. 814–822, Reykjavik, Iceland: PMLR.

Rasmussen CE, and Williams CKI (2005), Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning Series, London, England: MIT Press.

Reiss PT, Huang L, and Mennes M (2010), "Fast Function-on-Scalar Regression with Penalized Basis Expansions," The International Journal of Biostatistics, 6, 28. [PubMed: 21969982]

Rorden C, and Brett M (2000), "Stereotaxic Display of Brain Lesions," Behavioural Neurology, 12, 191–200. [PubMed: 11568431]

Shen X, Cox SR, Adams MJ, Howard DM, Lawrie SM, Ritchie SJ, Bastin ME, Deary IJ, McIntosh AM, and Whalley HC (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. [PubMed: 30093342]

Smith SM, and Nichols TE (2018), "Statistical Challenges in "big data" Human Neuroimaging," Neuron, 97, 263–268. [PubMed: 29346749]

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. [PubMed: 25826379]

Szaflarski JP, Altaye M, Rajagopal A, Eaton K, Meng X, Plante E, and Holland SK (2012), "A 10-year Longitudinal fMRI Study of Narrative Comprehension in Children and Adolescents," Neuroimage, 63, 1188–1195. [PubMed: 22951258]

Szaflarski JP, Schmithorst VJ, Altaye M, Byars AW, Ret J, Plante E, and Holland SK (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 KI Jr, Gebregziabher M, Kuchinsky SE, and Eckert MA (2012), "Multiple Imputation of Missing fMRI Data in Whole Brain Analysis," Neuroimage, 60, 1843–1855. [PubMed: 22500925]

Weiner KS, and Zilles K (2016), "The Anatomical and Functional Specialization of the Fusiform Gyrus," Neuropsychologia, 83, 48–62. [PubMed: 26119921]

Welling M, and Teh YW (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 WH (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, 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. [PubMed: 37008532]

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. [PubMed: 25435598]

Figure 1.
Incremental differences of BIOS, SBIOS0, and SBIOSimp.

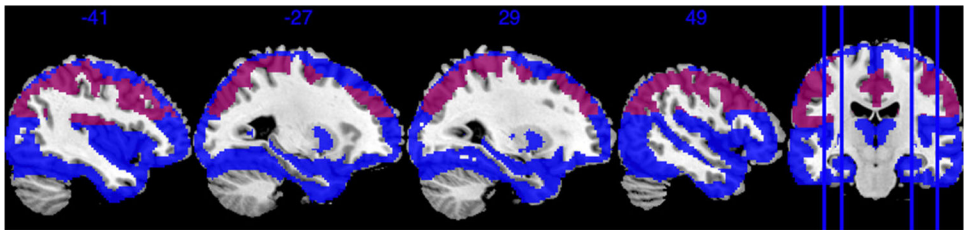
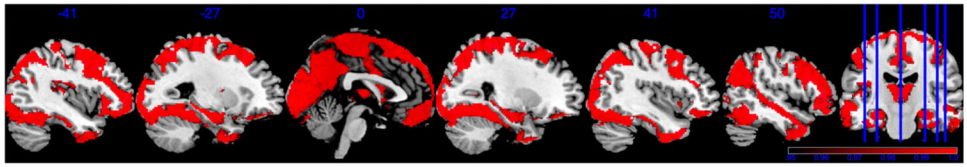
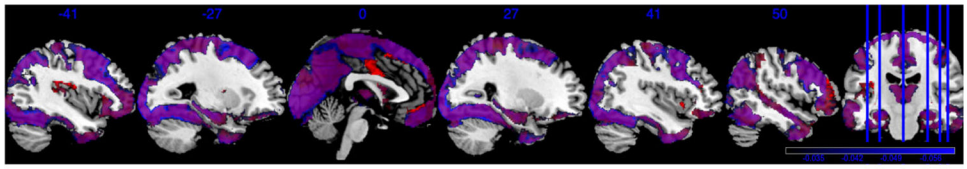


Figure 2.

Analysis mask using an observed proportion threshold of 0.5 and an intersection mask (completely observed data). The purple area indicates 100% inclusion; the blue area indicates the mask with an observed proportion between 0.5 and 1.0.



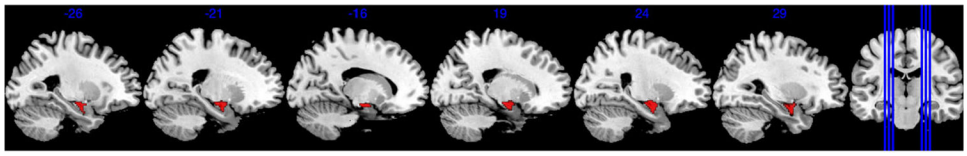
(a) Posterior inclusion probability (PIP). The color bar from black to red ranges in [0.95, 1]. Sagittal plane. The first two sagittal slices are in the left hemisphere.



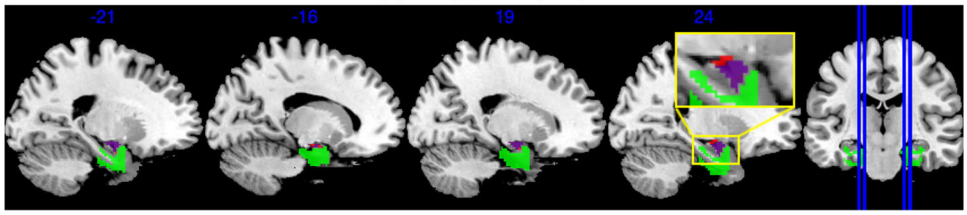
(b) Posterior mean of $\beta(s)\delta(s)$. The color bar from black to blue ranges from -0.03 to -0.06. The overlaying purple area is the mask of PIP greater than 0.95.

Figure 3.

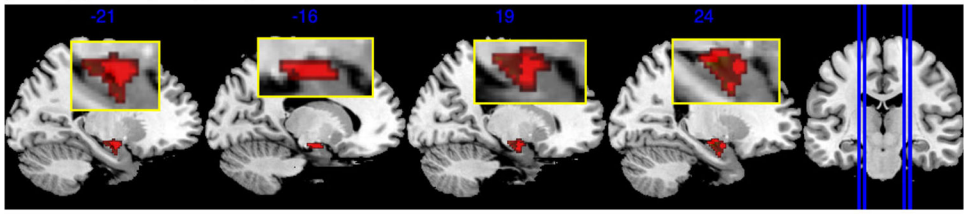
Illustration of age-related activation patterns using a grayscale brain background image (`ch2bet`, Holmes et al. 1998). Images are created using MRIcron (Rorden and Brett 2000).



(a) Amygdala (red).



(b) Purple area indicates voxels in the amygdala with PIP > 0.95. Green area is the *parahippocampal gyrus, anterior division*



(c) Yellow shaded area indicates voxels in the amygdala with at least 50% decline in the brain signal intensity for 10-year increase in age from 50.

Figure 4.
Illustrations on the amygdala region.

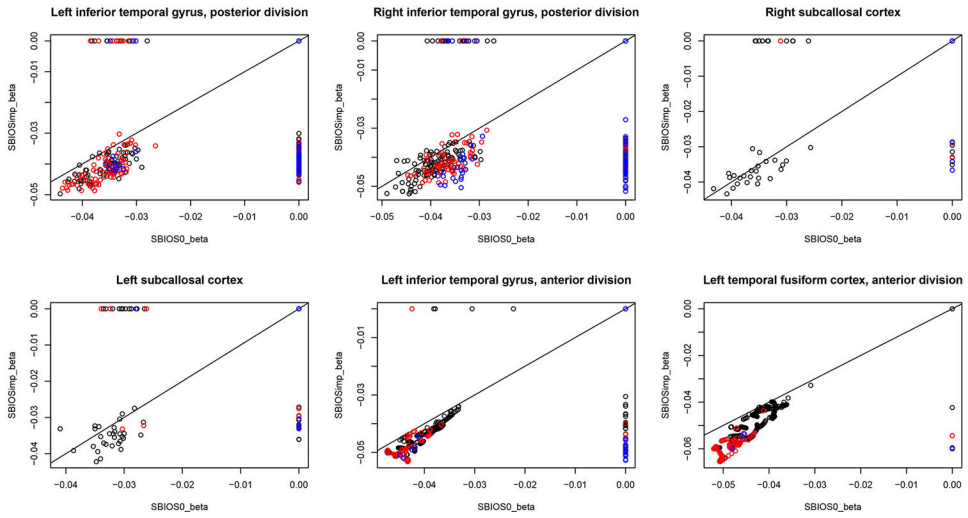
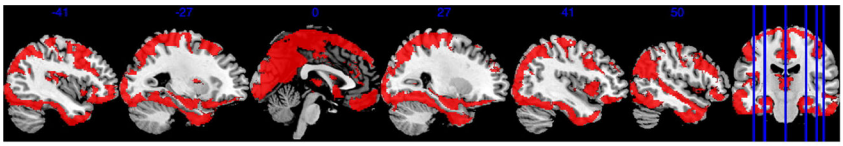
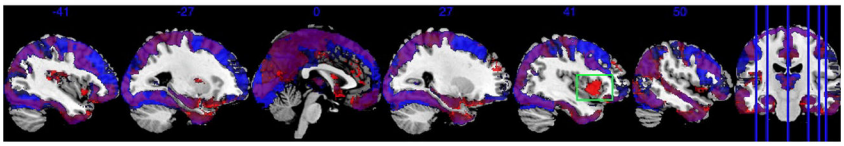


Figure 5.

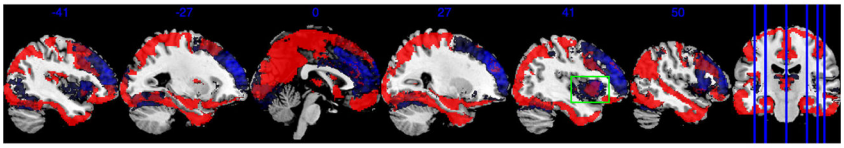
Scatterplot of the posterior mean of $\beta(s_j)/(\text{PIP}(s_j) \geq 0.95)$ based on SBIOS0 (x-axis) and SBIOSimp (y-axis) on six selected regions with high missingness. Blue dots indicate voxels with observed proportion $h(s_j) \in [0.5,0.7)$. Red dots indicate voxels with observed proportion $h(s_j) \in [0.7,0.9)$. Black dots indicate voxels with observed proportion $h(s_j) \in [0.9,1]$.



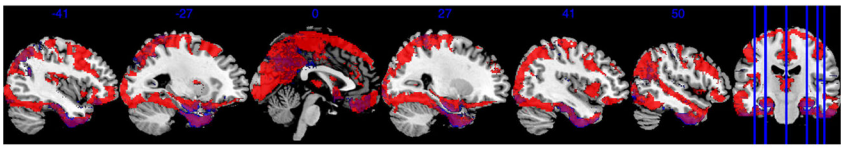
(a) Posterior inclusion probability (PIP). The color bar from black to red ranges in $[0.95, 1]$. Sagittal plane. The first two sagittal slices are in the left hemisphere.



(b) Posterior mean of $\beta(s)$. The color bar from black to blue ranges from -0.03 to -0.06. The overlaying purple or red area is the mask of PIP greater than 0.95.



(c) Posterior mean of $\tilde{\beta}(s)$ (positive). The color bar from black to blue ranges from 0.01 to 0.06. The overlaying purple or red area is the mask of PIP greater than 0.95.



(d) Posterior mean of $\tilde{\beta}(s)$ (negative). The color bar from black to blue ranges from -0.01 to -0.02. The overlaying purple or red area is the mask of PIP greater than 0.95.

Figure 6.

Illustration of results after applying a common $\delta(s)$ term on both the main effect of age and the interaction effect between age and gender, using a grayscale brain background image (`ch2bet`, Holmes et al. (1998)). Images are created using MRIcron (Rorden and Brett 2000). The highlighted green boxed area indicates one active region that has small effect size for $\beta(s)$ but large effect size for $\tilde{\beta}(s)$.

Table 1.

Top 10 regions ordered by Region Level Activation Rate (RLAR).

Region Name	Size	RLAR	Neg sum (count)	Median 50–60
Right intracalcarine cortex	634	1	−47.08(632)	−53.83
Right supracalcarine cortex	151	1	−8.38(151)	−52.15
Left temporal fusiform cortex, anterior division	301	1	−15.57(299)	−81.18
Left inferior frontal gyrus, pars triangularis	692	1	−48.86(686)	−56.38
Right inferior temporal gyrus, anterior division	314	1	−16.54(303)	−93.38
Left intracalcarine cortex	557	0.99	−40.29(547)	−51.38
Right temporal fusiform cortex, anterior division	276	0.99	−14.81(263)	−75.87
Left occipital pole	1977	0.99	−157.61(1911)	−57.67
Left hippocampus	218	0.98	−12.51(209)	−77.37
Left inferior temporal gyrus, anterior division	346	0.98	−15.26(316)	−106.83

NOTE: The 4th column reports the negative voxel effect size summed over each region, that is, $\sum_{j \in \mathcal{B}_j} \mathbb{E}\{\beta(s_j) \mid \beta(s_j) < 0\}$, and inside the bracket are the number of negative voxels. The last column is the percentage changes in the brain intensity when age increases from 50 to 60 (see Section S4.1). 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.

Table 2.

Proportion of active voxels identified by MUA and SBIOSimp for each range of observed proportions (OP).

OP	[0.5, 0.7)	[0.7, 0.9)	[0.9, 1]
MUA	0.15	0.39	0.62
SBIOSimp	0.27	0.46	0.62

Table 3.

Sensitivity analysis of SBIOSimp on UK Biobank data.

Region ID	Region Name	Cases 1&2	Case 3	Case 4	Case 5
24	(R) intracalcarine cortex	1.00	1.00	1.00	1.00
47	(R) supracalcarine cortex	1.00	1.00	1.00	1.00
85	(L) temporal fusiform cortex, AD	1.00	1.00	1.00	1.00
53	(L) inferior frontal gyrus, PT	1.00	1.00	1.00	1.00
14	(R) inferior temporal gyrus, AD	1.00	1.00	0.99	0.99
72	(L) intracalcarine cortex	0.99	0.99	0.99	0.99
37	(R) temporal fusiform cortex, AD	0.99	0.99	0.99	1.00
96	(L) occipital pole	0.99	0.99	0.99	0.99
105	(L) hippocampus	0.98	0.99	0.98	0.99
62	(L) inferior temporal gyrus, AD	0.98	0.98	0.99	0.99

NOTE: The RLAR for each region is reported. For the region names, (R) means right hemisphere, (L) means left hemisphere, AD means anterior division, PT means pars triangularis.

Table 4.

Region level results when applying δ to both age and the interaction of age and gender, on the same regions as reported in Table 1.

Region name	Size	RLAR	Neg Sum (count)	Neg/count
Right intracalcarine cortex	634	0.95	−39.38(521)	−0.076
Right supracalcarine cortex	151	1.00	−7.04(151)	−0.047
Left temporal fusiform cortex, anterior division	301	1.00	−15.05(301)	−0.050
Left inferior frontal gyrus, pars triangularis	692	0.99	−42.64(654)	−0.065
Right inferior temporal gyrus, anterior division	314	1.00	−14.77(312)	−0.047
Left intracalcarine cortex	557	0.99	−35.74(534)	−0.067
Right temporal fusiform cortex, anterior division	276	1.00	−13.84(275)	−0.050
Left occipital pole	1977	0.84	−120.70(1303)	−0.093
Left hippocampus	218	0.98	−11.99(211)	−0.057
Left inferior temporal gyrus, anterior division	346	1.00	−14.32(346)	−0.041

NOTE: See the caption of Table 1 for the interpretation of column names.