Bayesian Analysis (2020) **15**, Number 4, pp. 1229–1249

# High Dimensional Single-Index Bayesian Modeling of Brain Atrophy

Arkaprava Roy[*],[§] Subhashis Ghosal[†],[§] and Kingshuk Roy Choudhury[‡]
For The Alzheimer's Disease Neuroimaging Initiative[¶]

**Abstract.** We propose a model of brain atrophy as a function of high-dimensional genetic information and low-dimensional covariates such as gender, age, APOE gene, and disease status. A nonparametric single-index Bayesian model of high-dimension is proposed to model the relationship using B-spline series prior on the unknown functions and Dirichlet process scale mixture of centered normal prior to the distributions of the random effects. The posterior rate of contraction without the random effect is established for a fixed number of regions and time points with increasing sample size. We implement an efficient computation algorithm through a Hamiltonian Monte Carlo (HMC) algorithm. The performance of the proposed Bayesian method is compared with the corresponding least square estimator in the linear model with horseshoe prior, Least Absolute Shrinkage and Selection Operator (LASSO) and Smoothly Clipped Absolute Deviation (SCAD) penalization on the high-dimensional covariates. The proposed Bayesian method is applied to a dataset on volumes of brain regions recorded over multiple visits of 748 individuals using 620,901 SNPs and 6 other covariates for each individual, to identify factors associated with brain atrophy.

**Keywords:** ADNI, Bayesian, Genome-wide association study (GWAS), Hamiltonian Monte Carlo, high-dimensional data, single-index Model, spike-and-slab prior.

## 1 Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease that affects approximately 5.5 million people in the United States and about 30 million people worldwide. It is believed to have a prolonged preclinical phase initially characterized by the development of silent pathologic changes when patients appear to be clinically normal, followed by mild cognitive impairment (MCI) and then dementia (AD) (Petrella (2013)). Apart from its manifestation in the impairment of cognitive abilities, disease progression also produces a number of structural changes in the human brain, which includes the deposition of amyloid protein and the shrinkage or atrophy for certain regions of the

[*] Department of statistics, Duke University, aroy2@ncsu.edu

[†] Department of statistics, North Carolina State University, sghosal@ncsu.edu

[‡] Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, kingshuk.roy.choudhury@duke.edu

[§] Research is partially supported by NSF grant DMS-1510238.

[¶] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

https://doi.org/10.1214/19-BA1186

brain over time (Thompson et al. (2003)). Previous studies have shown that the rate of brain atrophy is significantly modulated by a number of factors, such as gender, age, baseline cognitive status and most markedly, allelic variants in the Apolipoprotein E (APOE) gene (Hostage et al. (2014)). In this paper, we examine if any other genes are also implicated in modulating the rate of brain atrophy along with examining effects of the low-dimensional covariate on the rate of atrophy using the data, collected by Alzheimer's Disease Neuroimaging Initiative (ADNI).

We model regional brain volume, which has been measured from magnetic resonance (MR) images using a segmentation procedure and recorded in the ADNI database. We collect this data directly from ADNI. We have volumetric measurements over six visits for thirteen disjoint brain regions and a total brain measure which is a summary measure of total brain parenchyma, including the Cerebral-Cortex, Cerebellum-Cortex, Thalamus-Proper, CaudatePutamen, Pallidum, Hippocampus, Amygdala, Accumbens-area, VentralDC, Cerebral-White-Matter, Cerebellum-White-Matter, and WM-hypointensities. These visits were roughly around six months apart. Thus the subjects were scanned roughly over three years. The rate of atrophy differs across different individuals and is assumed to be dependent on subject-specific covariates like genetic variations, gender, age, etc. In this paper, we propose a model to study the effects of these different covariates along with Alzheimer's disease state on atrophy in different brain regions. This analysis represents a technical challenge because the genomic data is high-dimensional and needs to be incorporated in a model for longitudinal progression of brain volumes measured in multiple parts of the brain in a non-parametric setup. A schematic of the regions, we studied here, are depicted in the Figure 1. This image is reproduced with permission from Ahveninen et al. (2012). The pre-processed brain volume data, used in this paper is obtained from the ADNI database. Hostage et al. (2014) had also used a similar set of regions of interest.

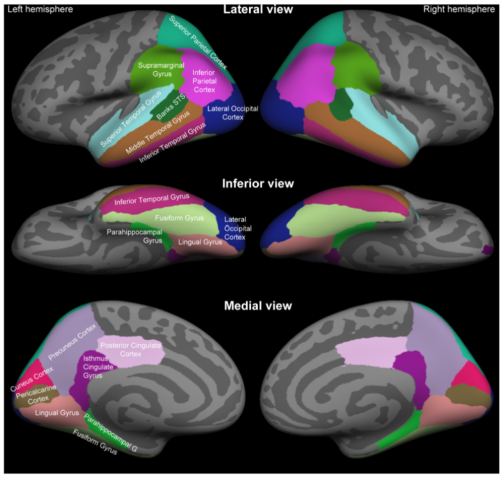

Figure 1: Anatomic parcellation of cortical surface from different angles showing brain regions used for analysis. (courtesy: Ahveninen et al. (2012).)

We consider two separate sets of unknown functions of covariates $X$ and $Z$ to model the volumetric measurements of the first visit and rates of changes for different regions. These functions have two inputs. The first $X$ consists of high-dimensional single-nucleotide polymorphism (SNP) of each individual, and the other $Z$ consists of low-dimensional covariates like gender, age, disease state, APOE gene status of each individual. The effect of these covariates on brain volume is modelled using a semi-parametric function of the form $\{a_{0,j}(X'\beta, Z'\eta) : j = 1, \ldots, 14\}$ for the volumetric measurements of the first visit and $\{a_{1,j}(X'\beta, Z'\eta) : j = 1, \ldots, 14\}$ for the rate of change in the $j$th region and the coefficients $\beta$ and $\eta$ are unit vectors of appropriate dimensions. A finite random series prior is put on the functions based on tensor products of B-splines with appropriate prior distribution on the coefficients. To address the issues associated with the high-dimensionality of $X$, the coefficient $\beta$ is assumed to be sparse. We reparametrize $\beta$ in polar coordinates to incorporate sparsity in its prior. Figure 1 of the Supplementary Materials (Roy et al., 2019) shows that the logarithm of the total brain measure changes non-linearly with time. This is the case for other brain regions as well. In addition to that, Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog) and Clinical Dementia Rating scale Sum of Boxes (CDR-SB) has been found to change quadratically with time (Lin et al., 2015). Thus, we incorporate the effect of time in the modeling by an increasing function which is estimated nonparametrically. Here also, we use a finite random series of B-splines with an appropriate prior on the coefficients to model this unknown function of time.

Apart from proposing a sophisticated model for studying brain atrophy, the proposed method develops a new estimation scheme for a general high-dimensional single-index model (high-dimensional SIM). Estimation for high-dimensional single-index model was previously addressed in Zhu and Zhu (2009), Yu and Ruppert (2002), Wang et al. (2012), Peng and Huang (2011), Radchenko (2015) and Luo and Ghosal (2016). All of them used the $\ell_1$-penalty and used optimization techniques to compute the estimates. In a Bayesian framework, Antoniadis et al. (2004) used the Fisher–von Mises prior to the directional vector. This cannot be easily modified for a high-dimensional covariate as then we shall need a prior which favors many zeros in the unit vector. Another paper addressing sparse Bayesian single-index model estimation is Alquier and Biau (2013). Even though their method is theoretically attractive, it is difficult to implement for high-dimensional covariate due to its high computational complexity. Wang (2009) developed a Bayesian method for the sparse single-index model using the reversible jump Markov chain Monte Carlo (MCMC) technique which is computationally expensive, especially in the high-dimensional setting. We introduce a new way of incorporating sparsity on a unit vector by a spike-and-slab prior via the polar form. The computation scheme is based on an efficient gradient-based Hamiltonian Monte Carlo (HMC) algorithm.

The rest of the paper is organized as follows. Section 2 discusses dataset and modeling in more detail. In Section 3, we describe the prior on different parameters of the model. Section 4 describes posterior computation in this setup. We study the posterior rate of contraction in the model in Section 5 under the asymptotic regime that the number of individuals goes to infinity but the number of time points where measurements are taken and the number of regions is fixed. We present a simulation study comparing the proposed Bayesian procedure with its linear counterpart in Section 6. The concentration of the posterior justifies the use of the proposed Bayesian procedure from a frequentist perspective. In Section 7, we present conclusions from the ADNI data on brain atrophy described above using our proposed method. Section 8, concludes the paper with some further remarks.

## 2 Data description and modeling

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to predict the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). In the ADNI dataset, the grey matter part of the brain is divided into thirteen disjoint regions. The volume of these regions and the summary measure of the whole brain are recorded over time for $n = 748$ individuals. The visits are scheduled after every 6 months over a span of 36 months. Not all of the participants turned up at each of those scheduled visits. There was no record of visiting after $30^{th}$ month for any of the participants. The volume data of $J = 14$ regions which include thirteen brain regions and the summary measure of the whole brain over $T_i$ set of time points which are original visit times in months divided by 36 to keep it bounded in [0, 1] for the $i$th individual, for $i = 1, \dots, 748$ is collected where $1 \leq |T_i| \leq 6$. The notation $|T_i|$ denotes cardinality of the set $T_i$. For some of the participants, the disease status changed during the span of this study. However, they were very small in number. We do not include them in this study.

In ADNI, the subjects were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA), yielding a set of 620,901 SNP and copy number variation (CNV) markers. The APOE gene has been the most significant gene in GWAS of Alzheimer's disease. The corresponding SNPs, rs429358 and rs7412, are not on the Human 610-Quad Bead-Chip. At the time of participant enrollment, APOE genotyping was performed and included in the ADNI database. The two SNPs (rs429358, rs7412) define the epsilon 2, 3, and 4 alleles and therefore were not genotyped using DNA extracted by Cogenics from a 3 mL aliquot of Ethylenediaminetetraacetic acid (EDTA) blood. These alleles are considered separately in the study.

Thus apart from the volumetric measurements, we also have high-dimensional SNP data and data on some other covariates for each individual. The other covariates are gender, disease state, age and allele 2 and 4 of the APOE gene. Except for the covariate age, all other low-dimensional covariates are categorical. To represent the categories, we use binary dummy variables. Since the disease status has three states—NC (no cognitive impairment), MCI (mild cognitive impairment) and AD (Alzheimer's disease), we consider two dummy variables $Z^{\mathrm{AD}}$ and $Z^{\mathrm{MCI}}$ respectively are the dummy indicator variables for MCI and AD, setting NC as the reference group. Similarly, the dummy variable $Z^{\mathrm{M}}$ indicating male gender is introduced setting females as the reference group. Also, we introduce $Z^{\mathrm{APOE},2}$, $Z^{\mathrm{APOE},4}$ standing for Alleles 2 and 4 for the two alleles

APOEallele2 and APOEallele4 together setting Allele 3 as a reference group for each of the two cases. We consider the age corresponding to the initial visit as a covariate as well. Let $Z = (Z^{\mathrm{MCI}}, Z^{\mathrm{AD}}, Z^{\mathrm{M}}, \mathrm{Age}, Z^{\mathrm{APOE},2}, Z^{\mathrm{APOE},4})$ stand for the whole vector of covariates. The continuous variable is Age.

With time, different brain regions change differently. We study the effects of different attributes to these changes. For every individual, the volume of a brain region on a particular visit should primarily depend on the volume of that region at the baseline visit and the rate of change of volume for that region with time. These rates of changes are region-specific as well as individuals-specific. Hence, it is logical to consider the baseline volume and the rate of change as functions on the subject and brain region. For simplicity, we do not assume any form of spatial dependence between measurements across brain regions. Thus we need two sets $\{a_{0,j}(\cdot,\cdot), a_{1,j}(\cdot,\cdot) : j = 1, \ldots, 14\}$ of functions for modeling volume at the initial visit and the rate of change for the $j$th region. These functions are unknown and are modeled nonparametrically. For nonparametric regression problems, single-index models provide a lot of flexibility in the estimation and interpretation of the results. Hence we adopt the bivariate single-index model with two inputs for high-dimensional and low-dimensional covariates separately for easy interpretation and computational efficiency. The effect of time is captured through an unknown increasing function $F_0(\cdot)$, which is bounded in $[0, 1]$. This is also modeled nonparametrically. The region-specific rate of change function $a_{1,j}(X_i'\beta, Z_i'\eta)$, which is multiplied with $F_0$, is dependent on subjects and regions. To make this function identifiable, we consider $F_0$, independent of subjects or regions.

Let $Y_{ij}(t)$ be the volume in the logarithmic scale of the $j$th brain region for the $i$th individual at the $t$th time point, $X_i$ is high-dimensional SNP expression of length $p$ for the $i$th individual, $t \in T_i$, $i = 1, \ldots, m$ and $j = 1, \ldots, 14$ and $Z_i$ is low-dimensional covariate of length $k$ for the $i$th individual. Then the data generating process can be represented through the following specification

$$\begin{aligned} Y_{ij}(t) &= F_{ij}(t) + \varepsilon_{ij}(t), \varepsilon_{ij}(t) \sim^{\mathrm{iid}} \mathrm{N}(0, \sigma^2), \\ F_{ij}(t) &= a_{0,j}(X_i'\beta, Z_i'\eta) - a_{1,j}(X_i'\beta, Z_i'\eta) F_0(t), \end{aligned} \tag{2.1}$$

where $a_{0,j}(\cdot,\cdot), a_{1,j}(\cdot,\cdot), F_0(\cdot)$ are all unknown continuous functions and N stands for a normal distribution and $\sim^{\mathrm{iid}}$ signifies that error $\varepsilon_{ij}(t)$s are independent across $i, j$ and $t$ and follows the distribution $\mathrm{N}(0, \sigma^2)$. The function $F_0(\cdot)$ is monotone increasing functions from $[0, 1]$ to $[0, 1]$ and $F_0(0) = 0$, $F_0(1) = 1$. The other two functions $a_{0,j}(\cdot,\cdot), a_{1,j}(\cdot,\cdot)$ maps from $[-1, 1]^2$ to $(-\infty, \infty)$. For identifiability of the functions along with the parameters $\beta$ and $\eta$, we assume that $\|\beta\| = 1$, $\|\eta\| = 1$; here $\|\cdot\|$ denotes $L_2$-norm of a vector. We also normalize the covariates for each individual i.e. $X_i$ and $Z_i$ for each $i$ such that $\|X_i\|$ and $\|Z_i\|$ are one. This is to ensure that $X_i'\beta$ and $Z_i'\eta$ are bounded between $[-1, 1]$ for each $i$. For nonparametric function estimation, bounded domain is important for uniform approximation using the basis expansion. The biggest challenge for estimation in this model is the high-dimensionality of $\beta$. To identify important SNPs from $X$, we need to perform variable selection. To do that, we propose a sparse estimation scheme for $\beta$. First we reparametrize the two unit vector $\beta = (\beta_1, \ldots, \beta_p)$ and $\eta = (\eta_1, \ldots, \eta_k)$ to their respective polar forms which allow us to work with Euclidean spaces. In this setup, only the directions of $\beta$ and $\eta$ are identifiable. Note that $\beta$ and $-\beta$ have the same directions. In the polar setup, for $s = 1, \ldots, p-1$, $\beta_s = \prod_{l=1}^{s-1} \sin\theta_l \cos\theta_s$, and $\beta_p = \prod_{l=1}^{p-1} \sin\theta_l$ where $\{\theta = (\theta_1, \ldots, \theta_{p-1})\}$ is the polar angle corresponding to the unit vector $\beta$. Here $\theta_s \in [0, \pi]$ for $s = 1, \ldots, (p-2)$ and $\theta_{p-1} \in [0, 2\pi]$. Similarly, let $\alpha$ be the polar angle corresponding to $\eta$. Then for $s \leq k-1$, $\eta_s = \prod_{l=1}^{s-1} \sin\alpha_l \cos\alpha_s$ and $\eta_k = \prod_{l=1}^{k-1} \sin\alpha_l$.

## 3 Prior specification

In the nonparametric Bayesian setup described above, we induce prior distributions on the smooth functions $a_{0,j}$ and $a_{1,j}$ in (2.1) through basis expansions in tensor products of B-splines and suitable prior for the corresponding coefficients. Given other parameters in this setup, a normal prior distribution on the coefficients of the tensor products of B-splines will lead to conjugacy and faster sampling via Gibbs sampling scheme. An inverse gamma prior on $\sigma^2$ is an obvious choice due to conjugacy and faster sampling. We also put a B-spline series prior on the smooth increasing function of time $F_0(\cdot)$. The coefficients for this function would be increasing in the index of the basis functions and lie in (0,1]. To put a prior on an increasing sequence, we introduce a set of latent variables of size equal to one less than the number of B-spline coefficients. Then the B-spline coefficients would be normalized the cumulative sum of those latent variables. Other two parameters $\beta$ and $\eta$ are reparametrized to their polar coordinate system. The parameter space of the polar angles will be a hyper-rectangle. It will be easier to put prior to the polar angles. To estimate using the sparsity of $\beta$, we need to carefully put a shrinkage prior on the polar angles. A polar angle of $\pi/2$ will ensure that the corresponding coordinate in the unit vector equals to zero. When there is sparsity in the unit vector, most of the polar angles will be $\pi/2$. Thus a spike-and-slab prior on the polar angle with a spike at $\pi/2$ should be able to capture sparsity in the corresponding unit vector. The last polar angle has spike both at the multiples of $\pi$ and $\pi/2$, due to the special structure of the last and the penultimate co-ordinates of a unit vector in the polar form. If it is a multiple of $\pi$, then the last coordinate is zero and if it is an odd multiple of $\pi/2$, then the penultimate coordinate is zero. The support of the last polar angle is $[0, 2\pi]$ and we need spikes at all multiples of $\pi$ and $\pi/2$. The prior is constructed to maintain that structure of spikes. Since only the directions of $\beta$ and $\eta$ are identifiable, the intercept and slope functions are modeled as even functions i.e. symmetric around zero.

Now we describe the prior in details. Let $\lceil x \rceil$ denote the lowest integer greater than or equal to $x$.

(i) Intercept and slope functions: Let $a_{\nu,j}(x,y) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \lambda_{\nu,j,mm'} B_m(x) B_{m'}(y)$, $\nu = 0, 1$, with $\lambda_{\nu,j,mm'} = \lambda_{\nu,j,(K-m)m'}$, $\lambda_{\nu,j,mm'} = \lambda_{\nu,j,m(K-m')}$ for $\nu = 0, 1$, and $\lambda_{\nu,j,mm'} \sim \mathrm{N}(0, a^2)$, $1 \leq m, m' \leq \lceil K/2 \rceil$, for some chosen $a > 0$.

(ii) The function of time: Let $F_0(x) = \sum_{l=1}^{K'} \lambda_l B_l(x)$, with $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{K'} = 1$. We put a prior on $(\lambda_2, \ldots, \lambda_{K'-1})$ by reparameterizing as $\lambda_{l+1} = (\sum_{i=1}^{l} \delta_i)/(\sum_{i=1}^{K'-1} \delta_i)$, and putting the prior $\delta_i \sim \mathrm{Un}(0, 1)$ for $l = 1, \ldots, K'-1$, where Un stands for the uniform distribution.

(iii) Error variance: We put $\sigma^{-2} \sim \mathrm{Ga}(d_1, d_2)$, where Ga stands for the gamma distribution.

(iv) Polar angles $\alpha$ of $\eta$: We let $\alpha_r \sim \mathrm{Un}(0, \pi)$, $r = 1, \ldots, (k-2)$, and $\alpha_{k-1} \sim \mathrm{Un}(0, 2\pi)$, independently.

(v) Polar angles $\theta$ of $\beta$: We put a spike-and-slab prior on the polar angles that has spike at $\pi/2$ for the first polar angle and all the multiples of $\pi/2$ for the last polar angle. Then the spike distribution would look like Figure 2. The spike-and-slab prior for $\theta_i$, $i \leq (p-2)$, has density given by

$$(1-\gamma_i)\frac{\Gamma(M_1+M_2)}{\Gamma(M_1)\Gamma(M_2)}\left(\frac{\min(\theta_i, \pi-\theta_i)}{\pi/2}\right)^{M_2}\left(1-\frac{\min(\theta_r, \pi-\theta_i)}{\pi/2}\right)^{M_1}+\gamma_i\frac{1}{\pi},$$

for $0 < \theta_i < \pi$ and the distribution of $\theta_{p-1}$ is given by

$$\begin{aligned}\theta_{p-1} \quad \sim \quad & (1-\gamma_i)\frac{1}{8}[\mathrm{Be}(M_1, M_2)_{[0,\pi/4]} + \mathrm{Be}(M_2, M_1)_{[\pi/4,\pi/2]} \\ & +\mathrm{Be}(M_1, M_2)_{[\pi/2,3\pi/4]} + \mathrm{Be}(M_2, M_1)_{[3\pi/4,\pi]} + \mathrm{Be}(M_1, M_2)_{[\pi,5\pi/4]} \\ & +\mathrm{Be}(M_2, M_1)_{[5\pi/4,3\pi/2]} + \mathrm{Be}(M_1, M_2)_{[3\pi/2,7\pi/4]} + \mathrm{Be}(M_2, M_1)_{[7\pi/4,2\pi]}] \\ & +\gamma_i \mathrm{Un}(0, 2\pi);\end{aligned}$$

here $\mathrm{Be}(M_1, M_2)_{[a,b]}$ denotes the beta distribution with shape parameters $M_1$ and $M_2$, supported within the interval $[a, b]$, and $M_1 < 1 \leq M_2$. The indicator variable $\gamma \sim \mathrm{Ber}(q)$.

The spike distribution on $\theta$ looks like Figure 2. The first plot is for the first $(p-2)$ angles and the second plot is for the last polar angle.

*Model selection*: Polar angles with a posterior probability of selection in the model more than 0.5 are considered in the model.

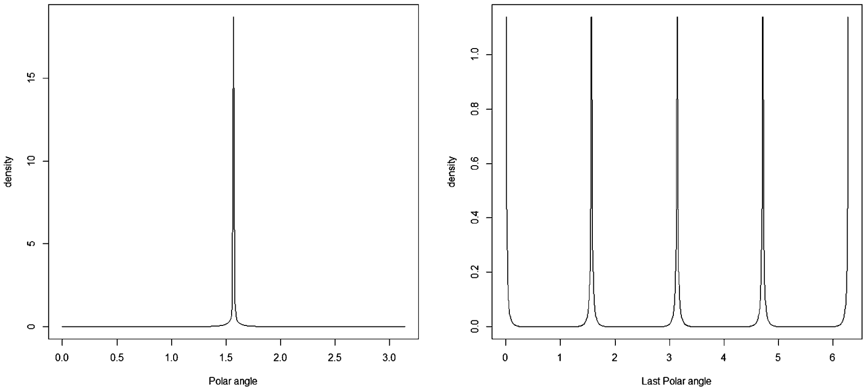

Figure 2: Spike distribution with $M_1 = 0.01$ and $M_2 = 10$.

# 4 Computation

The conditional log-likelihood is given by

$$
\begin{aligned}
C - \sum_{ijt} \frac{1}{2\sigma^2} \big(Y_{ij}(t) - \sum_{m=1}^{K}\sum_{m'=1}^{K} \lambda_{0,j,mm'} B_m(\sum_{s=1}^{p-1} X_{is} \prod_{l=1}^{s-1} \sin\theta_l \cos\theta_s + X_{ip} \prod_{l=1}^{p-1} \sin\theta_l) \\
\times B_{m'}(\sum_{s=1}^{k-1} Z_{is} \prod_{l=1}^{s-1} \sin\alpha_l \cos\alpha_s + Z_{ik} \prod_{l=1}^{k-1} \sin\alpha_l) \\
+ \sum_{m=1}^{K}\sum_{m'=1}^{K} \lambda_{1,j,mm'} B_m(\sum_{s=1}^{p-1} X_{is} \prod_{l=1}^{s-1} \sin\theta_l \cos\theta_s + X_{ip} \prod_{l=1}^{p-1} \sin\theta_l) \\
\times B_{m'}(\sum_{s=1}^{k-1} Z_{is} \prod_{l=1}^{s-1} \sin\alpha_l \cos\alpha_s + Z_{ik} \prod_{l=1}^{k-1} \sin\alpha_l) \sum_{m=1}^{K-1} \frac{\sum_{i=1}^{m} \delta_i}{\sum_{i=1}^{K'-1} \delta_i} B_{m+1}(t)\big)^2 \\
- \sum_{m,m',j} \frac{\lambda_{0,j,mm'}^2 + \lambda_{1,j,mm'}^2}{2a^2} \\
+ \sum_{l=1}^{p-1} \log\big((1-\gamma_l)\theta_l'^{\,M_1-1}(1-\theta_l')^{M_2-1} \frac{\Gamma(M_1+M_2)}{\Gamma(M_1)\Gamma(M_2)} + \gamma_l\big) \\
- (J \sum_i T_i/2 + d_1 - 1) \log\sigma^2 - d_2/\sigma^2,
\end{aligned}
$$

where $C$ involves only the hyperparameters $a, M_2, M_1, K, d_1, d_2$ and the observations but not the parameters of the model.

All B-spline coefficients and $\sigma$ are updated using the conjugacy structure. Using expressions for the derivative of B-splines (De Boor, 2001), it is possible to calculate the derivatives of the log-likelihood with respect to the polar angles and $\delta$ of $F_0(t)$. Thus, we implement an efficient gradient-based MCMC algorithm using Hamiltonian Monte Carlo algorithm, described in Neal (2011). Since these parameters have bounded support, the candidates for a Metropolis step are reflected back if they cross the boundaries. For example if a candidate $\theta_i^c$ of $\theta_i$ is more than $\pi$, then it is re-adjusted as $\theta_i^c = \pi - (\theta_i^c - \pi)$. A similar adjustment is done if $\theta_i^c$ becomes smaller than zero as well. Similar strategies are adopted while updating $\delta$'s as well. All the posterior updates are discussed in the supplementary materials.

The parameters $K$ and $K'$ are chosen by minimizing the Bayesian Information Criterion (BIC) after fitting the model over a grid of a number of B-spline basis functions for randomly generated 50 different choices of $\beta$ and $\eta$. We generate 50 different choices for $\beta$ and $\eta$ from the prior and then fit the non-linear model for different choices of a number of B-spline basis functions within the range [7, 20]. After taking the average over all 10 BIC values for each case, we choose the number that has the least BIC value as the optimal value. The convergence of the MCMC chain is diagnosed using trace plots.

# 5 Large-sample properties

In this section, we examine the large-sample properties of the proposed Bayesian procedure for the model (2.1). We have observations for fixed $J$ number of regions and $T$ many time points. We show posterior consistency in the asymptotic regime of increasing sample size and increasing dimension $p$ of the SNPs. For sake of generality of the method, $K$ is given a prior with probability mass function of the form $\Pi(K = k) = b_1 \exp[-b_2 k^2 (\log k)^{b_3}]$, with $b_1, b_3 > 0$ and $0 \leq b_3 \leq 1$. The functions $B_m$, $m = 1, \dots, K$, stand for B-spline basis. Also, $K'$ is given a prior with probability mass function of the form $\Pi(K' = k) = b_1' \exp[-b_2' k (\log k)^{b_3'}]$ with $b_1', b_2' > 0$ and $0 \leq b_3' \leq 1$. Here also, we denote B-spline basis functions as $B_m$. Note that either geometric or Poisson distribution (respectively $b_3' = 0$ and 1) can be chosen as prior on $K'$, the number of terms to be used in the B-spline series for the growth function $F_0$. For $K$, the square, which is the number of terms in the tensor product series representation, can have a geometric or Poisson-like tail. In our computation of the model, we are not using these priors on $K$ and $K'$ i.e. the number of B-spline basis functions as it will require reversible jump MCMC strategy which is computationally expensive.

We study the posterior contraction rate with respect to the empirical $\ell_2$-distance on the regression function, which is defined as follows. For two sets of parameters $(F, a, \beta, \eta)$ and $(F^*, a^*, \beta^*, \eta^*)$, the empirical $\ell_2$-distance is given by

$$\begin{aligned} & d^2((F, a, \beta, \eta), (F^*, a^*, \beta^*, \eta^*)) \\ & = \frac{1}{J \sum_{i=1}^n T_i} \sum_{j=1}^J \sum_{i=1}^n \sum_{t=1}^{T_i} |a_{0,j}(X_i'\beta, Z_i'\eta) - a_{1,j}(X_i'\beta, Z_i'\eta) F(t) \\ & \qquad - a^*_{0,j}(X_i'\beta^*, Z_i{}''\eta^*) + a^*_{1,j}(X_i'\beta^*, Z_i'\eta^*) F^*(t)|^2, \end{aligned}$$

and $a = (a_{0,j}, a_{1,j} : j = 1, \dots, J)$, $a^* = (a^*_{0,j}, a^*_{1,j} : j = 1, \dots, J)$.

Since $\beta$ is a high-dimensional parameter, sensible estimation is possible only if it has sparsity, which must be picked up by the prior. In the setting of a spike-and-slab prior for polar coordinates described in Section 3, we need to ensure sufficient concentration near $\pi/2$ by choosing a large value of the second parameter $M_2$ in the beta spike distribution (depending on the sample size $n$ and the dimension $p$) and a small value of the probability of slab $\gamma$. More precisely, we choose $M_1 \leq 1$ fixed, $M_2 > \sqrt{np} \log p$ and $\gamma = o(p^{-1})$. Then the contraction rate will be determined by the smoothness of the underlying true functions $a_{0,0}, a_{0,1}$ and $F_{0,0}$, the sparsity $s_0$ of true high-dimensional regression coefficient $\beta_0$ and mildly on the parameter $b_3, b_3'$ in the prior distribution for the number of basis elements $K, K'$ used in the B-spline bases, as shown by the following result.

**Theorem 1.** *Assume that the true function $F_{0,0}$ belongs to the Hölder class of regularity level $\iota'$ on $[0, 1]$ and the true coefficient functions $(a_{0,0,j}, a_{1,0,j} :, j = 1, \dots, J)$ are Hölder smooth function of regularity level $\iota$ on $[-1, 1] \times [-1, 1]$. Let the true regression coefficient $\beta_0$ for the high-dimensional covariate $X$ have $s_0$ non-zero co-ordinates. Then* *the posterior contraction rate with respect to the distance $d$ is*

$$\max\left\{n^{-\iota/(2\iota+2)}(\log n)^{\iota/(2\iota+2)+(1-b_3)/2}, n^{-\iota'/(2\iota'+1)}(\log n)^{\iota'/(2\iota'+1)+1-b_3'}, \sqrt{\frac{s_0\log p}{n}}\right\}.$$

In the above result, since the observation points are not dense over the domain, posterior contraction on the function is based on its distance with the true function only at the observation points.

The proof of the theorem uses the general theory of posterior contraction (see Ghosal and van der Vaart (2017)) for independent non-identically distributed observations and some estimates for finite random series based on B-splines. The proof is given in the supplementary materials part.

# 6 Simulation

We compare our method with some common penalization methods based on the following simplified linear model

$$\log Y_{ij}(t) = X_i'\beta + Z_i'\eta + \gamma_{1,j} - (X_i'\beta + Z_i'\eta + \gamma_{2,j})t + \varepsilon_{ij}(t), \tag{6.1}$$

$t = 1, \dots, T_i$, $i = 1, \dots, n$, $j = 1, \dots, 13$. In the above model, $\beta$ is a sparse vector and all other parameters are unpenalized. The performance of these methods is compared based on MSE values on a test set under the scenarios the linear model is correct and the linear model is false.

We generate a high-dimensional binary matrix $X$ and a low-dimensional covariate matrix $Z$. The true data matrices of the real data are in the support of the scheme of sampling the high-dimensional matrix as well as the low-dimensional matrix.

*Data generation for the non-linear case:*

- Generate a data matrix $X$ with elements coming from Bernoulli distribution with success probability of the $i$th row as $p_i$.
- Generate $p_i$, $i = 1, \dots, n$, from the standard uniform distribution.
- Generate all the elements of the matrix $Z$ from $\mathrm{N}(0, 1)$.
- Generate the sparse vector $\beta_0$ with 5% elements non-zero. Positions for non-zero elements are chosen first at random by sampling $p/20$ elements from total $p$ positions, where $p$ is the length of $\beta_0$. The non zero elements are generated from mixture distribution of two normals $\mathrm{N}(2, 1)$ and $\mathrm{N}(-1, 1)$.
- Set the value of $\eta_0$ to $(1, -2, 4.3, 10, -8)$.
- Normalize each row of $X$ and $Z$ along with $\beta_0$ are $\eta_0$ to the unit norm.

We let the true functions be

$$a_{0,0,j}(x,y) = 2((j/13)x)^3 + 2((1-j/13)y)^3,$$
$$a_{0,1,j}(x,y) = 2\exp((j/13)y) + 2\exp((1-j/13)x),$$

$j = 1,\dots,13$, and $F_{0,0}(t) = t^2$. After generating the true functional values, the data $Y_{ij}(t)$ is generated from $\mathrm{N}(a_{0,j}(X_i'\beta_0, Z_i'\eta_0) - a_{1,j}(X_i'\beta_0, Z_i'\eta_0)F_0(t), 1)$.

*Data generation for the linear case:*

In this case, the true model is the one given in (6.1). All the steps for generating $X$, $Z$, $\beta_0$ and $\eta_0$ are as given above. The coefficients $\gamma_{1,j}$ and $\gamma_{2,j}$, $j = 1,\dots,13$, are generated from $\mathrm{N}(0,1)$. After generating the design matrix, the data $Y_{ij}(t)$ is generated from $\mathrm{N}(X_i\beta_0 + Z_i\eta_0 + \gamma_{1,j} - (X_i\beta_0 + Z_i\eta_0 + \gamma_{2,j})t, 1)$.

For the data generation, we set the error standard deviation to $\sigma_0 = 1$. We compare the prediction MSEs across all the methods. We split the whole data into two equal parts for training and testing. We compare the performance of our method with the LASSO (Tibshirani (1996)), the SCAD (Fan and Li (2001)) and the horseshoe method (Carvalho et al., 2010) on testing dataset based estimates of the parameters from the training dataset. We use the R package `glmnet` for LASSO and `ncvreg` for SCAD. We develop and use an HMC sampler for the horseshoe prior. For sample sizes 200, 500 and 1000, we gather the mean squared error (MSE) values for non-linear and linear cases. We use half of the sample for training and the remaining half for testing. Among other parameters, we consider thirteen regions in total, five-time points and vary the value of $p$ as 5000, 10000 and 20000. We set $M_1 = 0.1, M_2 = 10$ and tune $q$ to ensure a good acceptance rate and desired model size (sum of the $\gamma_i$) across MCMC samples. We consider maintaining the desired model size between 20 and 30 at each step of the MCMC iterations. The results are summarized below for 50 replications and 3000 post-burn samples after burn-in 1000 samples. The number of basis functions for spline is different across different sample sizes. To fit our model we vary the number of B-spline basis functions as 8 for $n = 200$, 11 for $n = 500$, and 14 for $n = 1000$. These numbers are chosen according to the strategy described in the Section 4. Let $I_0$ be the set of indices such that $\{i : \beta_{0i} \neq 0\}$. We have in total of 26 non-zero variables. From the posterior samples of $\gamma$, we can identify the top 26 selected variables. For other methods, we select the variables with the highest 26 absolute values in $\hat{\beta}$ and ignore the ones with very low magnitudes relative to others. Let $\hat{X}$ denote the final selected set of variables for different cases and let $X_{I_0}$ be the true set of variables. To examine the performance of variable selection, we report the maximum canonical correlation between $\hat{X}$ and $X_{I_0}$ for each case in the bracket.

From Table 1, we infer that the performance of the proposed Bayesian method based on the high-dimensional single-index model is always much better than the LASSO, the SCAD, and the horseshoe method for the non-linear case. For the linear case in Table 2, it is competitive with linearity based methods like the LASSO, the SCAD, and the horseshoe method. This is natural as the LASSO, the SCAD or the horseshoe method use more precise modeling information which the semiparametric methods cannot use. However, in terms of maximum canonical correlations, our model outperforms in the non-linear case and remains extremely competitive in the linear case.

| Total<br>sample size | Dimension<br>of $\beta$ ($p$) | SIM<br>MSE | LASSO<br>MSE | SCAD<br>MSE | Horseshoe<br>MSE |
|---|---|---|---|---|---|
| 200 | 5000 | 3.33 (0.84) | 6.93 (0.37) | 7.32 (0.71) | 6.24 (0.60) |
| 500 | 5000 | 3.43 (0.75) | 6.11 (0.32) | 6.03 (0.69) | 7.46 (0.61) |
| 1000 | 5000 | 3.27 (0.76) | 6.36 (0.18) | 7.08 (0.69) | 6.89 (0.60) |
| 200 | 10000 | 3.42 (0.84) | 6.32 (0.12) | 7.16 (0.69) | 8.33 (0.59) |
| 500 | 10000 | 3.09 (0.74) | 6.69 (0.34) | 7.80 (0.66) | 7.18 (0.58) |
| 1000 | 10000 | 3.25 (0.77) | 6.01 (0.35) | 7.13 (0.67) | 7.84 (0.58) |
| 200 | 20000 | 3.40 (0.83) | 6.88 (0.07) | 7.75 (0.70) | 5.24 (0.61) |
| 500 | 20000 | 3.31 (0.74) | 6.54 (0.34) | 7.20 (0.67) | 7.54 (0.60) |
| 1000 | 20000 | 3.15 (0.75) | 6.12 (0.25) | 7.03 (0.70) | 7.86 (0.55) |

Table 1: Comparison of the proposed high-dimensional single-index model with the LASSO, the SCAD, and the horseshoe method in terms of MSE and the maximum canonical correlation between a selected set of variables and true set of variables in the bracket for the non-linear case.

| Total<br>sample size | Dimension<br>of $\beta$ ($p$) | SIM<br>MSE | LASSO<br>MSE | SCAD<br>MSE | Horseshoe<br>MSE |
|---|---|---|---|---|---|
| 200 | 5000 | 1.67 (0.92) | 1.02 (0.59) | 1.01 (0.86) | 1.01 (0.72) |
| 500 | 5000 | 1.48 (0.71) | 1.01 (0.43) | 1.02 (0.74) | 1.01 (0.66) |
| 1000 | 5000 | 1.61 (0.71) | 1.00 (0.49) | 1.02 (0.70) | 1.00 (0.66) |
| 200 | 10000 | 1.61 (0.78) | 1.02 (0.39) | 1.01 (0.76) | 1.00 (0.78) |
| 500 | 10000 | 1.48 (0.65) | 1.01 (0.38) | 1.02 (0.66) | 1.02 (0.58) |
| 1000 | 10000 | 1.65 (0.66) | 1.01 (0.40) | 1.02 (0.64) | 1.01 (0.59) |
| 200 | 20000 | 1.35 (0.79) | 1.03 (0.22) | 1.01 (0.73) | 1.01 (0.65) |
| 500 | 20000 | 1.24 (0.72) | 1.02 (0.35) | 1.01 (0.73) | 1.00 (0.66) |
| 1000 | 20000 | 1.31 (0.70) | 1.01 (0.45) | 1.02 (0.65) | 1.01 (0.60) |

Table 2: Comparison of the proposed high-dimensional single-index model with the LASSO, the SCAD, and the horseshoe method in terms of MSE and the maximum canonical correlation between a selected set of variables and true set of variables in the bracket for the linear case.

# 7 Real-data analysis

## 7.1 Modification of the model for real data application

### Incorporating random effects and regions wise varying effect

As the data are longitudinal, it is reasonable to add a subject specific random effect ($\tau_i$) in the model. We also vary the coefficient $\eta$ of the low-dimensional covariates region-wise following the linear model of Hostage et al. (2014). However, we keep the high-dimensional coefficient $\beta$ fixed for all the regions because of the high computational cost. Thus, the selected SNPs are responsible for the change in all the brain regions. The new modified model will then become

$$Y_{ij}(t) = F_{ij}(t) + \tau_i + \varepsilon_{ij}(t), \quad \varepsilon_{ij}(t) \sim \mathrm{N}(0, \sigma^2),$$

$$F_{ij}(t) = a_{0,j}(X_i'\beta, Z_i'\eta_j) - a_{1,j}(X_i'\beta, Z_i'\eta_j)F_0(t), \tag{7.1}$$

$t = 1, \ldots, T_i$ with $1 \leq T_i \leq 6$, $j = 1, \ldots, 14$, $i = 1, \ldots, 748$.

*Prior on the random effects:*

We put a Dirichlet process scale mixture of normal prior to the random effect distribution.

### Region-wise varying effect with no SNP

To compare the nonlinear model with the linear model of Hostage et al. (2014), we also fit the following model without the SNPs:

$$\begin{aligned} Y_{ij}(t) &= F_{ij}(t) + \tau_i + \varepsilon_{ij}(t), \quad \varepsilon_{ij}(t) \sim \mathrm{N}(0, \sigma^2), \\ F_{ij}(t) &= a_{0,j}(Z_i'\eta_j) - a_{1,j}(Z_i'\eta_j)F_0(t), \end{aligned} \tag{7.2}$$

$t = 1, \ldots, T_i$ with $1 \leq T_i \leq 6$, $j = 1, \ldots, 14$, $i = 1, \ldots, 748$.

### Corresponding linear model

We compare the performance of our method based on the above non-linear model with the following linear model based on Hostage et al. (2014),

$$Y_{ij}(t) = H_{ij}(t) + \tau_i + \varepsilon_{ij}(t), \quad \varepsilon_{ij}(t) \sim \mathrm{N}(0, \sigma^2),$$

where $t = 1, \ldots, T_i$ with $1 \leq T_i \leq 6$, $j = 1, \ldots, 14$, $i = 1, \ldots, 748$, and

$$\begin{aligned} H_{ij}(t) =& \varrho_{j0} + \varrho^0_{j,\mathrm{M}} Z_{i,\mathrm{M}} + \varrho^0_{j,\mathrm{AD}} Z_{i,\mathrm{AD}} + \varrho^0_{j,\mathrm{NC}} Z_{i,\mathrm{NC}} + \varrho^0_{j,\mathrm{Allele4}} Z_{i,\mathrm{Allele4}} \\ &+ \varrho^0_{j,\mathrm{Allele2}} Z_{i,\mathrm{Allele2}} + \varrho^0_{j,\mathrm{Age}} Z_{i,\mathrm{Age}} + \varrho^0_{j,\mathrm{AD,Allele2}} Z_{i,\mathrm{AD}} Z_{i,\mathrm{Allele2}} \\ &+ \varrho^0_{j,\mathrm{AD,Allele4}} Z_{i,\mathrm{AD}} Z_{i,\mathrm{Allele4}} + \varrho^0_{j,\mathrm{NC,Allele2}} Z_{i,\mathrm{NC}} Z_{i,\mathrm{Allele2}} \\ &+ \varrho^0_{j,\mathrm{NC,Allele4}} Z_{i,\mathrm{NC}} Z_{i,\mathrm{Allele4}} - \big[\varrho_{j1} + \varrho^1_{j,\mathrm{M}} Z_{i,\mathrm{M}} + \varrho^1_{j,\mathrm{AD}} Z_{i,\mathrm{AD}} + \varrho^1_{j,\mathrm{NC}} Z_{i,\mathrm{NC}} \\ &+ \varrho^1_{j,\mathrm{Allele4}} Z_{i,\mathrm{Allele4}} + \varrho^1_{j,\mathrm{Allele2}} Z_{i,\mathrm{Allele2}} + \varrho^1_{j,\mathrm{Age}} Z_{i,\mathrm{Age}} \\ &+ \varrho^1_{j,\mathrm{AD,Allele2}} Z_{i,\mathrm{AD}} Z_{i,\mathrm{Allele2}} + \varrho^1_{j,\mathrm{AD,Allele4}} Z_{i,\mathrm{AD}} Z_{i,\mathrm{Allele4}} \\ &+ \varrho^1_{j,\mathrm{NC,Allele2}} Z_{i,\mathrm{NC}} Z_{i,\mathrm{Allele2}} + \varrho^1_{j,\mathrm{NC,Allele4}} Z_{i,\mathrm{NC}} Z_{i,\mathrm{Allele4}}\big] t. \end{aligned} \tag{7.3}$$

We have the volumetric measurement data for the total thirteen brain regions along with the summary measure of the whole brain over time for 748 individuals. For each individual, the covariate information is summarized in Table 3. The reference group for our analysis is a female individual with average age and no cognitive impairment.

We first fit the model in (7.2) and the following linear model in (7.3) in accordance with Hostage et al. (2014) with the same set of covariates and interactions between APOE and disease states. Then we compare the prediction MSE. The prediction error

| | Female | Male |
|---|---|---|
| Number of NC participants | 99 | 114 |
| Number of AD participants | 84 | 94 |
| Number of MCI participants | 122 | 235 |
| Number of participants with APOEgene allele2 | 29 | 29 |
| Number of participants with APOEgene allele4 | 150 | 223 |
| Average age | 73.51 | 74.60 |
| Standard deviation of Age | 6.67 | 6.80 |

Table 3: Demographic table to summarize data in terms of number of no cognitive impaired (NC), Alzheimer's disease (AD) and mildly cognitive impaired (MCI) participants along with the age across the two gender groups male and female.

gives us predictive performance and fitted relative MSE helps to judge the reliability of inference. We consider a basis consisting of 17 B-spline functions for univariate and $17^2$ basis functions for bivariate cases. The estimates are based on 5000 post-burn MCMC samples after burning-in the first 1000 samples.

To compare the fitted models, we calculate the prediction error in each model. To calculate the prediction error, we divide the whole dataset into training (Tr) and testing (Te) sets. We use stratified sampling using each subject-region pair as stratum so that training will have all the individuals that belong to the testing set. This is important for prediction with a random effect in the model. The formula for prediction error will be $|\mathrm{Te}|^{-1}\sum_{(i,j,t)\in\mathrm{Te}}(Y_{ij}(t)-\hat{Y}_{ij}(t))^2$; here $|\mathrm{Te}|$ denotes total number of elements in the test set Te. The linear model gives the prediction error of 3.83 whereas that in our non-linear model hugely improves to 0.06. The model in (7.1) with SNPs improves the prediction error to 0.45 which is about 25% improvement. Region-wise prediction errors for the models in (7.1) and (7.3) are provided in Table 1 and 2 of the Supplementary Materials respectively. Figure 3 shows the estimated $F_0(t)$ function for the model in (7.1). The estimated effect of time is indeed non-linear. This suggests a non-linear change in the logarithm of volume with time for a given individual and a brain region.

After selecting the significant genes, we calculate the Bayesian information criterion (BIC) of models leaving out one of the low-dimensional covariates every time with all the genes to compare the significance. If BIC for the model leaving out covariate A is higher than the model leaving out covariate B, then covariate A is more significant than covariate B. The table below gives an ordered list of the significance of low-dimensional covariates for different regions in Tables 4. In Table 5, we show the estimates from the linear model in (7.3) for the whole brain.

We map the significant SNPs from our analysis to the corresponding genes using the R package `rsnps`. We tune the parameter $\gamma$ in the model to select the 20 most significant SNPs. Among those, we could map 11 of those to some genes. The significant genes from our analysis are mentioned in Section 8 along with some previous studies that found the corresponding gene significant for the AD and/or cerebral atrophy.

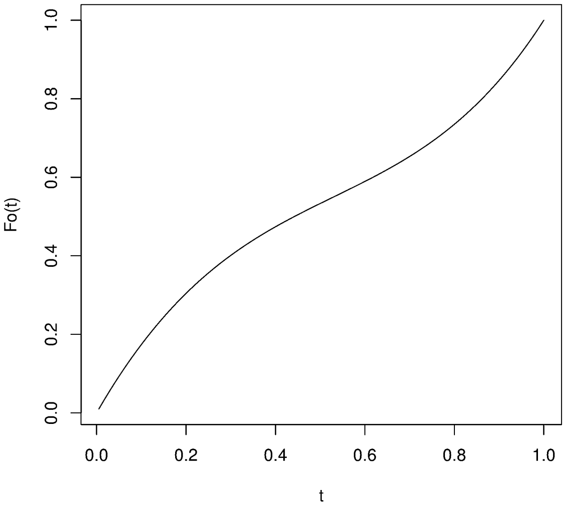

Figure 3: Estimated $F_0(t)$ function for model in (7.1).

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Total Brain | Age | APOE-allele4 | APOE-allele2 | Gender | MCI | AD |
| Ventricles | MCI | APOE-allele4 | Gender | AD | APOE-allele2 | Age |
| Left-Hippocampus | APOE-allele4 | Gender | MCI | APOE-allele2 | AD | Age |
| Right-Hippocampus | MCI | APOE-allele4 | APOE-allele2 | Gender | AD | Age |
| LINFLATVEN | APOE-allele4 | Gender | AD | MCI | APOE-allele2 | Age |
| RINFLATVEN | APOE-allele4 | AD | MCI | Gender | APOE-allele2 | Age |
| Left Medial-Temporal | MCI | Age | APOE-allele4 | APOE-allele2 | Gender | AD |
| Right Medial-Temporal | Gender | APOE-allele4 | APOE-allele2 | Age | MCI | AD |
| Left Inferior-Temporal | APOE-allele4 | APOE-allele2 | MCI | Age | Gender | AD |
| Right Inferior-Temporal | APOE-allele4 | MCI | APOE-allele2 | AD | Age | Gender |
| Left Fusiform | Gender | MCI | Age | APOE-allele4 | APOE-allele2 | AD |
| Right Fusiform | APOE-allele4 | Age | MCI | APOE-allele2 | Gender | AD |
| Left Entorhin | APOE-allele4 | Gender | MCI | Age | AD | APOE-allele2 |
| Right Entorhin | APOE-allele4 | Gender | Age | MCI | AD | APOE-allele2 |

Table 4: low-dimensional covariates of the model in (7.2) with selected genes for different brain regions in their order of significance, 1 being the most significant.

| | Value | Std.Error | Degree of Freedom | t-value | p-value |
|---|---|---|---|---|---|
| time | 0.013 | 0.001 | 2155 | 13.208 | 0.000 |
| APOEallele4:time | 0.000 | 0.001 | 2155 | 0.047 | 0.962 |
| APOEallele2:time | −0.001 | 0.002 | 2155 | −0.249 | 0.804 |
| Gender:time | 0.003 | 0.001 | 2155 | 3.077 | 0.002 |
| MCI:time | 0.006 | 0.001 | 2155 | 4.811 | 0.000 |
| AD:time | 0.014 | 0.002 | 2155 | 6.969 | 0.000 |
| Age:time | −0.002 | 0.000 | 2155 | −4.663 | 0.000 |
| APOEallele4:MCI:time | 0.004 | 0.002 | 2155 | 2.689 | 0.007 |
| APOEallele4:AD:time | 0.004 | 0.002 | 2155 | 2.105 | 0.035 |
| APOEallele2:MCI:time | 0.001 | 0.003 | 2155 | 0.292 | 0.770 |
| APOEallele2:AD:time | −0.003 | 0.006 | 2155 | −0.530 | 0.596 |

Table 5: Estimates of covariates for Total Brain for slope from linear model.

## 8 Conclusions and discussion

We fit a bivariate single-index model to capture the volumetric change of different cortical regions in the human brain. There are both high and low-dimensional covariates as input in the unknown functions determining the initial configuration and the rate of change of different regions. To tackle the high-dimensional covariate within a single-index model, we provide a new technique of assigning a sparse prior in this paper and propose an efficient MCMC scheme using a Hamiltonian Monte Carlo sampling. We present a result on posterior consistency of the resulting procedure. An 'R' package is attached to this paper as supplementary material (https://github.com/royarkaprava/High-Dimensional-Single-index-model).

In our results on the real dataset, we find that allele 4 of the APOE gene is always among the top three significant covariates for almost all the cases in Table 4. The fact that allele 4 of the APOE gene is significant was established in Hostage et al. (2014). They used a linear model, similar to the model in (7.3). Allele 4 is not found to be significant for the linear case in Table 5 as well as for several other regions in the linear case. However it is always ranked among the top three most significant covariates for the non-linear model. Thus, for this dataset, the linear model is not appropriate. We identify 11 significant genes. There are some previous studies that also noted the significant genes from our analysis as possible candidates for the AD and/or cerebral atrophy. The genes along with associated future study citing that gene in connection with AD and/or cerebral atrophy are mentioned here SLC6A1 (cerebellum) (Carvill et al., 2015), KCNIP4 (Himes et al., 2013), ADGRL3 (Orsini et al., 2016), SORBS2 (Zhang et al., 2016; Lee et al., 2014; Niceta et al., 2015), LPAR3 (Yung et al., 2015), SHROOM3 (Dickson et al., 2015; Freudenberg-Hua et al., 2016), SORCS3 (Breiderhoff et al., 2013; Lane et al., 2012), NPY2R (Lin et al., 2010; Schriemer et al., 2016), CWF19L2 (Lin et al., 2013), PALLD (Nho et al., 2015) and KCNMA1 (Burns et al., 2011; Tabarki et al., 2016). In particular, SLC6A1 has been found to affect cerebellum (Carvill et al., 2015); ADGRL3 affects hippocampus, the prefrontal cortex, and the striatum (Orsini et al., 2016); LPAR3 affects central and peripheral nervous tissues (Yung et al., 2015);

SORCS3 and PALLD affect hippocampus (Breiderhoff et al., 2013; Nho et al., 2015); KCNMA1 affects olfactory bulb, cortex, basal ganglia, hippocampus, thalamus, cerebellum, vestibular nuclei, and spinal cord (Tabarki et al., 2016). Apart from this, Figure 3 suggests that with time logarithm of volume changes non-linearly for a given individual and a brain region. The estimated function closely resembles a cubic polynomial in shape, among the class of polynomial functions.

We have kept the low-dimensional covariates fixed with time. One interesting future direction would be to modify the model to incorporate time-varying covariates as well. For example, the disease status of some of the participants changed during the span of this study. Although they were very small in number and we ignored them from this study, it would be useful to consider them as well, using a more elaborate model. Also, we have not included the Mini-Mental State Exam (MMSE) scores for our analysis. It will be interesting to study the effect of that covariate as well. Our proposed model limits us to choose a unique set of genes across all the regions. We can modify to select separate sets of genes for different parts of the brain at the expense of additional computational burden. This would give us more insights into the interdependence between genes and different parts of the brain. In our computation of the model, we are not using the priors mentioned in Section 5 on $K$ and $K'$ i.e. the number of B-spline basis functions as it will require reversible jump MCMC strategy which is computationally expensive. The model will be richer if these priors can be incorporated.

A package to fit the high-dimensional single index model, as well as a linear regression model with Dirichlet-Laplace and horseshoe priors using the HMC algorithm, is given in `https://github.com/royarkaprava/High-Dimensional-Single-index-model`.

## Supplementary Material

Supplementary Materials of High-dimensional single-index Bayesian modeling of brain atrophy (DOI: 10.1214/19-BA1186SUPP; .pdf).

## References

Ahveninen, J., Jääskeläinen, I. P., Belliveau, J. W., Hämäläinen, M., Lin, F., and Raij, T. (2012). "Dissociable influences of auditory object vs. spatial attention on visual system oscillatory activity." *Public Library of Science One*, 7(6): e38511. 1230

Alquier, P. and Biau, G. (2013). "Sparse single-index model." *Journal of Machine Learning Research*, 14: 243–280. MR3033331. 1231

Antoniadis, A., Grégoire, G., and McKeague, I. W. (2004). "Bayesian estimation In single-index models." *Statistica Sinica*, 14: 1147–1164. MR2126345. 1231

Breiderhoff, T., Christiansen, G. B., Pallesen, L. T., Vaegter, C., Nykjaer, A., Holm, M. M., Glerup, S., and Willnow, T. E. (2013). "Sortilin-related receptor SORCS3 is a postsynaptic modulator of synaptic depression and fear extinction." *Public Library of Science one*, 8(9): e75006. 1244, 1245

Burke, J. V. (2014). https://sites.math.washington.edu/~burke/crs/408/lectures/L3-Multivariable-Calc-Review.pdf.

Burns, L., Minster, R., Demirci, F., Barmada, M., Ganguli, M., Lopez, O., DeKosky, S., and Kamboh, M. (2011). "Replication study of genome-wide associated SNPs with late-onset Alzheimer's disease." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 156(4): 507–512. 1244

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1239

Carvill, G. L., McMahon, J. M., Schneider, A., Zemel, M., Myers, C. T., Saykally, J., Nguyen, J., Robbiano, A., Zara, F., and Specchio, N. (2015). "Mutations in the GABA transporter SLC6A1 cause epilepsy with myoclonic-atonic seizures." *The American Journal of Human Genetics*, 96(5): 808–815. 1244

De Boor, C. (2001). "A practical guide to splines, revised Edition." Vol. 27 of Applied Mathematical Sciences. *Mechanical Sciences, year*. MR1900298. 1236

Dickson, H. M., Wilbur, A., Reinke, A. A., Young, M. A., and Vojtek, A. B. (2015). "Targeted inhibition of the Shroom3–Rho kinase protein–protein interaction circumvents Nogo66 to promote axon outgrowth." *BMC neuroscience*, 16(1): 34. 1244

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its Oracle properties." *Journal of the American Statistical Association*, 96(1). MR1946581. doi: https://doi.org/10.1198/016214501753382273. 1239

Freudenberg-Hua, Y., Li, W., Abhyankar, A., Vacic, V., Cortes, V., Ben-Avraham, D., Koppel, J., Greenwald, B., Germer, S., and Consortium, T.-G. (2016). "Differential burden of rare protein truncating variants in Alzheimer's disease patients compared to centenarians." *Human molecular genetics*, 25(14): 3096–3105. 1244

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge. MR3587782. doi: https://doi.org/10.1017/9781139029834. 1238

Himes, B. E., Sheppard, K., Berndt, A., Leme, A. S., Myers, R. A., Gignoux, C. R., Levin, A. M., Gauderman, W. J., Yang, J. J., and Mathias, R. A. (2013). "Integration of mouse and human genome-wide association data identifies KCNIP4 as an asthma gene." *Public Library of Science one*, 8(2): e56179. 1244

Hostage, C. A., Choudhury, K. R., Doraiswamy, P. M., and Petrella, J. R. (2014). "Mapping the effect of the Apolipoprotein E Genotype on 4-Year Atrophy Rates in an Alzheimer Disease–related Brain Network." *Radiology*, 271(1). 1230, 1240, 1241, 1244

Jones, G. L. (2008). http://users.stat.umn.edu/~galin/icsprar.pdf.

Lane, R. F., St George-Hyslop, P., Hempstead, B. L., Small, S. A., Strittmatter, S. M., and Gandy, S. (2012). "Vps10 family proteins and the retromer complex in aging-related neurodegeneration and diabetes." *Journal of Neuroscience*, 32(41): 14080–14086. 1244

Lee, J. H., Cheng, R., Vardarajan, B. N., Lantigua, R. A., Reyes-Dumeyer, D., Ortmann, W., Graham, R., Bhangale, T., Behrens, T., and Medrano, M. (2014). "SORBS2, SH3RF3, and NPHP1 modify age at onset in carriers of the G206A mutation in PSEN1 with familial Alzheimer's disease." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 10(4): P632. 1244

Lin, J., Li, X., Yuan, F., Lin, L., Cook, C. L., Rao, C. V., and Lei, Z. (2010). "Genetic ablation of luteinizing hormone receptor improves the amyloid pathology in a mouse model of Alzheimer disease." *Journal of Neuropathology & Experimental Neurology*, 69(3): 253–261. 1244

Lin, K. A., Choudhury, K. R., Rathakrishnan, B. G., Marks, D. M., Petrella, J. R., Doraiswamy, P. M., Initiative, A. D. N., et al. (2015). "Marked gender differences in progression of mild cognitive impairment over 8 years." *Alzheimer's & dementia: translational research & clinical interventions*, 1(2): 103–110. 1231

Lin, P.-I., Kuo, P.-H., Chen, C.-H., Wu, J.-Y., Gau, S. S., Wu, Y.-Y., and Liu, S.-K. (2013). "Runs of homozygosity associated with speech delay in autism in a taiwanese han population: evidence for the recessive model." *Public Library of Science one*, 8(8): e72056. 1244

Luo, S. and Ghosal, S. (2016). "Forward selection and estimation in high dimensional single index models." *Stat Methodology*, 33: 172–179. MR3582782. doi: `https://doi.org/10.1016/j.stamet.2016.09.002`. 1231

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov Chain Monte Carlo*, 2(11): 2. MR2858447. 1236

Nho, K., Kim, S., Risacher, S. L., Ramanan, V. K., Shen, L., Foroud, T. M., Gibbons, L. E., Crane, P. K., Weiner, M. W., and Green, R. C. (2015). "Genome-wide rare variant analysis identifies candidate genes significantly associated with composite scores for memory." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 11(7): 251–252. 1244, 1245

Niceta, M., Stellacci, E., Gripp, K. W., Zampino, G., Kousi, M., Anselmi, M., Traversa, A., Ciolfi, A., Stabley, D., and Bruselles, A. (2015). "Mutations impairing GSK3-mediated MAF phosphorylation cause cataract, deafness, intellectual disability, seizures, and a Down syndrome-like facies." *The American Journal of Human Genetics*, 96(5): 816–825. 1244

Orsini, C. A., Setlow, B., DeJesus, M., Galaviz, S., Loesch, K., Ioerger, T., and Wallis, D. (2016). "Behavioral and transcriptomic profiling of mice null for Lphn3, a gene implicated in ADHD and addiction." *Molecular genetics & genomic medicine*, 4(3): 322–343. 1244

Peng, H. and Huang, T. (2011). "Penalized least squares for single index models." *Journal of Statistical Planning and Inference*, 141: 1362–1379. MR2747907. doi: `https://doi.org/10.1016/j.jspi.2010.10.003`. 1231

Petrella, J. (2013). "Neuroimaging and the search for a cure for Alzheimer disease." *Radiology*, 269: 671–691. 1229

Radchenko, P. (2015). "High dimensional single index models." *Journal of Multivariate Analysis*, 139: 266–282. MR3349492. doi: https://doi.org/10.1016/j.jmva.2015.02.007. 1231

Roy, A., Ghosal, S., and Choudhury, K. R. For The Alzheimer's Disease Neuroimaging Initiative (2019). "Supplementary Materials of High-dimensional single-index Bayesian modeling of brain atrophy." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1186SUPP. 1231

Schriemer, D., Sribudiani, Y., IJpma, A., Natarajan, D., MacKenzie, K. C., Metzger, M., Binder, E., Burns, A. J., Thapar, N., and Hofstra, R. M. (2016). "Regulators of gene expression in Enteric Neural Crest Cells are putative Hirschsprung disease genes." *Developmental biology*, 416(1): 255–265. 1244

Tabarki, B., AlMajhad, N., AlHashem, A., Shaheen, R., and Alkuraya, F. S. (2016). "Homozygous KCNMA1 mutation as a cause of cerebellar atrophy, developmental delay and seizures." *Human genetics*, 135(11): 1295–1298. 1244, 1245

Thompson, P., Hayashi, K., and deZubicaray G. (2003). "Dynamics of gray matter loss in Alzheimer's disease." *Journal of Neuroscience*, 23: 994–1005. 1230

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society B*, 58: 267–288. MR1379242. 1239

Wang, H. (2009). "Bayesian estimation and variable selection for single index models." *Computational Statistics and Data Analysis*, 53: 2617–2627. MR2665912. doi: https://doi.org/10.1016/j.csda.2008.12.010. 1231

Wang, T., Xu, P., and Zhu, L. (2012). "Non-convex penalized estimation in high-dimensional models with single-index structure." *The Journal of Multivariate Analysis*, 109: 221–235. MR2922865. doi: https://doi.org/10.1016/j.jmva.2012.03.009. 1231

Yu, Y. and Ruppert, D. (2002). "Penalized spline estimation for partially linear single index models." *Journal of American Statistical Association*, 97: 1042–1054. MR1951258. doi: https://doi.org/10.1198/016214502388618861. 1231

Yung, Y. C., Stoddard, N. C., Mirendil, H., and Chun, J. (2015). "Lysophosphatidic acid signaling in the nervous system." *Neuron*, 85(4): 669–682. 1244

Zhang, Q., Gao, X., Li, C., Feliciano, C., Wang, D., Zhou, D., Mei, Y., Monteiro, P., Anand, M., and Itohara, S. (2016). "Impaired dendritic development and memory in Sorbs2 knock-out mice." *Journal of Neuroscience*, 36(7): 2247–2260. 1244

Zhu, L. and Zhu, L. (2009). "Nonconcave penalized inverse regression in single-index models with high dimensional predictors." *The Journal of Multivariate Analysis*, 100(5): 862–875. MR2498719. doi: https://doi.org/10.1016/j.jmva.2008.09.003. 1231

## Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer's Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

We are grateful to the anonymous reviewers and the editor for their valuable comments that have greatly helped to improve the manuscript.