*The Annals of Applied Statistics*
2024, Vol. 18, No. 1, 704–728
https://doi.org/10.1214/23-AOAS1808

# A PARTIALLY FUNCTIONAL LINEAR REGRESSION FRAMEWORK FOR INTEGRATING GENETIC, IMAGING, AND CLINICAL DATA

BY TING LI[1,a], YANG YU[2,b], J. S. MARRON[2,c] AND HONGTU ZHU[3,d]

[1]*School of Statistics and Management, Shanghai University of Finance and Economics,* [a]*tingli@mail.shufe.edu.cn*
[2]*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill,* [b]*yangyu.unc@gmail.com,* [c]*marron@unc.edu*
[3]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [d]*htzhu@email.unc.edu*

This paper is motivated by the joint analysis of genetic, imaging, and clinical (GIC) data collected in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. We propose a partially functional linear regression (PFLR) framework to map high-dimensional GIC-related pathways for Alzheimer's disease (AD). We develop a joint model selection and estimation procedure by embedding imaging data in the reproducing kernel Hilbert space and imposing the $\ell_0$ penalty for the coefficients of genetic variables. We apply the proposed method to the ADNI dataset to identify important features from tens of thousands of genetic polymorphisms (reduced from millions using a preprocessing step) and study the effects of a certain set of informative genetic variants and the baseline hippocampus surface on 13 future cognitive scores. We also explore the shared and distinct heritability patterns of these cognitive scores. Analysis results suggest that both the hippocampal and genetic data have heterogeneous effects on different scores, with the trend that the value of both hippocampi are negatively associated with the severity of cognition deficits. Polygenic effects are observed for all the thirteen cognitive scores. The well-known APOE4 genotype only explains a small part of the cognitive function. Shared genetic etiology exists; however, greater genetic heterogeneity exists within disease classifications after accounting for the baseline diagnosis status. These analyses are useful in further investigation of functional mechanisms for AD progression.

## 1. Introduction.

Alzheimer's disease (AD) is a chronic neurodegenerative condition that leads to the degeneration of brain cells and a decline in cognitive, behavioral, and social skills. It is characterized by cognitive impairment, with significant variability in clinical presentation, as well as the burden and distribution of pathology among patients. This clinicopathologic heterogeneity presents both challenges and opportunities for conducting systematic and biomarker-based studies to enhance our understanding of AD biology, diagnosis, and management (Duong et al. (2022), Veitch et al. (2021)). AD has intricate pathophysiological mechanisms that are not yet fully understood. However, advances in biomarker identification, including genetic and imaging data, may aid in the early detection of individuals at risk for AD before the onset of symptoms.

The primary aim of this study is to use genetic, imaging, and clinical (GIC) variables from the ADNI study (Mueller et al. (2005)) to map the biological pathways of AD-related phenotypes of interest (e.g., cognition, intelligence, and progression status) (Jack et al. (2010, 2013), Duong et al. (2022), Veitch et al. (2021), Yu et al. (2022)). This may provide insights into the biological processes of brain development, healthy aging, and disease progression. For instance, integrating GIC variables is of great interest to elucidate the environmental, social, and genetic etiologies of intelligence and to delineate the foundations of intelligence

Received April 2023; revised August 2023.
*Key words and phrases.* Clinical, genetics, imaging, nonasymptotic error bounds, partially functional linear regression, sparsity.

differences in brain structure and function (Deary, Cox and Hill (2022)). Moreover, many brain-related disorders, including AD, are often caused by a combination of multiple genetic and environmental factors while being endpoints of abnormal brain structure and function (Knutson, Deng and Pan (2020), Shen and Thompson (2020)). A thorough understanding of such neurobiological pathways may lead to the identification of possible hundreds of risk genes, environmental risk factors, and brain structure and function abnormalities that underlie brain disorders. Once these risk genes and factors as well as brain abnormalities have been identified, it may be possible to detect them early enough to make a real difference in outcome and to develop related treatments, ultimately preventing the onset of brain-related disorders and reducing their severity.

To numerically map GIC-related pathways, we extract cognitive scores to quantify behavioral deficits, genetic covariates, demographic covariates at baseline, and brain imaging data from the ADNI study. As the hippocampus is particularly susceptible to AD pathology, it has become a major focus in AD research (Braak and Braak (1998)). We characterize the exposure of interest, hippocampal shape, using left/right hippocampal morphometry surface data represented as a 100 × 150 matrix. A detailed data description can be found in Section 2. Our focus lies in understanding how hippocampal shape and genetics can predict future cognitive deficits in Alzheimer's research. The unique data structure of these GIC variables presents new challenges for mapping the GIC pathway. First, conventional statistical tools, designed for scalar exposure, are not suitable for high-dimensional hippocampal imaging measures. Second, the dimension of the genetic covariates greatly exceeds the sample size. There is an urgent need for an effective statistical method capable of utilizing the hippocampal surface data and the ultrahigh-dimensional genetic data to map the GIC pathway.

To statistically map GIC-related pathways, we propose a high-dimensional partially functional linear regression (PFLR) model,

$$Y_i = \alpha + X_i^T \beta + \int_{\mathcal{T}} Z_i(t)\xi(t)\,dt + \epsilon_i \quad \text{for } i = 1, \dots, n. \tag{1}$$

Here $Y_i$ represents a continuous phenotype of interest, $X_i \in \mathcal{X}$ is a $p \times 1$ vector of genetic and environmental variables, and $Z_i(t) \in L_2(\mathcal{T})$ is an imaging (or functional) predictor over a compact set $\mathcal{T}$. The intercept term is denoted by $\alpha$, $\beta$ is a $p \times 1$ vector of coefficients, and $\xi(t)$ is an unknown slope function in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, while $\epsilon_i$ denotes measurement errors. We consider the scenario where the dimension of $\beta$ is either comparable to or much larger than the sample size $n$ and $\xi(t)$ is an infinite-dimensional function. Our goal is to make statistical inferences on $\beta$ and $\{\xi(t) : t \in \mathcal{T}\}$. As shown in Figure 1, there are genetic and clinical confounders that affect both hippocampal shape and behavioral deficits (Selkoe and Hardy (2016)). The classical high-dimensional linear model only considers genetic data. However, by including imaging exposure $\xi(t)$ in model (1), we can achieve two important implications for the ADNI dataset. First, the model can quantify

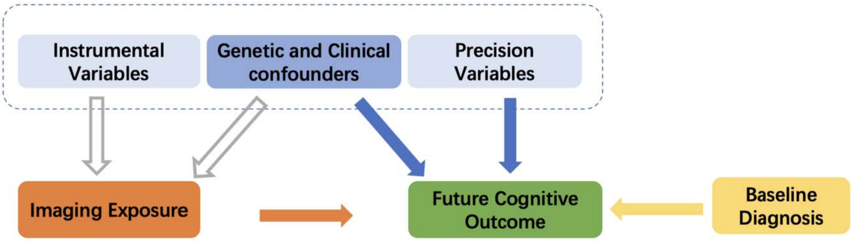


FIG. 1. *Directed acyclic graph showing potential relationships between the genetic data, the imaging data, and the future outcome. The four solid arrows denote the associations of interest.*

the direct effects of the confounders by controlling for the imaging exposure. Second, it can investigate the influence of the imaging exposure while controlling for the confounders and preserving the structure of the imaging data.

There is limited literature on partially functional linear regression (PFLR) with high-dimensional scalar covariates, with only a few notable exceptions. Kong et al. (2016) investigated PFLR in high dimensions, allowing the dimension of scalar covariates to diverge with the sample size, $n$. Yao, Sue-Chee and Wang (2017) developed a regularized partially functional quantile regression model, while also permitting the number of scalar predictors to increase with the sample size. Ma et al. (2019) focused on the partial functional quantile regression model in ultrahigh dimensions, with a diverging number of scalar predictors. All three of the aforementioned methods involve three steps: first, representing the functional predictors using their leading functional principal components (FPCs), second, reducing PFLR to a standard high-dimensional linear regression model, and third, selecting important features through the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)). As a result, existing approaches are heavily dependent on the success of the functional principal component analysis (FPCA) approach (Wang, Chiou and Müller (2016)). They may not be appropriate if the functional parameter cannot be effectively represented by the leading principals of the functional covariates (Yuan and Cai (2010)). Moreover, the truncation parameter in the FPCA changes discretely, which may result in imprecise control over model complexity, as noted in Ramsay and Silverman (2005).

Our PFLR is closely related to the growing literature on imaging genetics analysis, which has significantly expanded over the past decade and can be broadly divided into two major categories of statistical methods. The first category encompasses various statistical methods (Ge et al. (2015), Huang et al. (2017), Lin et al. (2014), Nathoo, Kong and Zhu (2019), Shen and Thompson (2020)) used in imaging genome-wide association studies (IGWAS) on imaging phenotypes. Thousands of brain-related genetic loci have been identified for regional grey matter volumes, white matter (WM) microstructure, and functional connectivity, with notable genetic overlaps observed with AD and other brain-related disorders (Elliott et al. (2018), Smith et al. (2021), Thompson et al. (2020), Zhao et al. (2022), Zhao et al. (2021), Zhao et al. (2019)). Two open resource knowledge portals for imaging genetics include the Oxford BIG40 (https://open.win.ox.ac.uk/ukbiobank/big40/) and BIG-KP (https://bigkp.org/). The second category involves various statistical methods for mapping biological pathways that link genetics and imaging data to brain-related disorders and examining their joint effects. This area remains challenging and has not been systematically studied (Zhu, Li and Zhao (2023)). Most existing methods initially extract features from the imaging data and focus on the effects of the obtained features and genetic data (Cruciani et al. (2022), Dukart, Sambataro and Bertolino (2016), Knutson, Deng and Pan (2020), Ossenkoppele et al. (2021)). However, this approach neglects the rich smoothness information present in the imaging data.

In this paper we focus on the high-dimensional PFLR (1), develop an estimation method for model selection and estimation, investigate the theoretical properties of both functional and scalar estimators, and apply the proposed method to analyze the ADNI dataset. We employ the RKHS framework (Yuan and Cai (2010), Cai and Yuan (2012), Li and Zhu (2020)) and impose the roughness penalty on the functional coefficient. Furthermore, we impose the $\ell_0$ penalty on the scalar predictors because the $\ell_0$ penalty function is often preferred among penalty functions, as it directly penalizes the cardinality of a model and seeks the most parsimonious model explaining the data. However, it is nonconvex, and solving an exact $\ell_0$-penalized nonconvex optimization problem involves exhaustive combinatorial best subset search, which is NP-hard and computationally challenging. We modify the computational algorithm in Huang et al. (2018) to address this difficulty and accommodate the functional predictor. Meanwhile, we adapt the test statistic in Li and Zhu (2020) to assess the significance of the functional variable. The implementation R code, along with its documentation, is available as an online supplement (https://github.com/BIG-S2/PFLR_RKHS_code). Numerically, the proposed method is rigorously tested on simulated data. We also provide theoretical properties of the estimators, including error bounds, asymptotic normality of the estimates of the nonzero scalar coefficients, and the null limit distribution of the test statistic designed to test the nullity of the functional variable.

We apply PFLR to the ADNI dataset and conduct a thorough association analysis between genetics, hippocampus, and cognitive deficit. Unlike existing analyses targeting one or several cognitive measures, the proposed method examines the joint effects of genetics and hippocampus on 13 cognitive variables observed 12 months after baseline measurements. These variables measure different aspects of cognitive function and explore the shared and distinct heritability patterns of the 13 cognitive scores. We also investigate the effect of baseline diagnosis information on future cognitive outcomes, denoted by the rightmost arrow in Figure 1. Analysis results suggest that both hippocampal and genetic data have heterogeneous effects on different scores. Generally, the value of both hippocampi are negatively associated with the severity of cognitive impairments. Polygenic effects are observed for all 13 cognitive scores, and shared genetic etiology exists. The strong genetic influence is only partially attributed to the well-known APOE4 genotype, and the baseline diagnosis status explains a larger part of cognitive function. There also exists a strong shared genetic effect beyond the effect of the APOE4 gene. However, greater genetic heterogeneity exists within disease classifications after accounting for the baseline diagnosis status. These analyses are useful for further investigation of functional mechanisms for AD progression.

The rest of this paper is organized as follows. Section 2 includes a detailed data and problem description. Section 3 describes our estimation procedure. Section 4 presents Monte Carlo simulation studies to assess the finite sample performance of the proposed method. Section 5 provides a detailed data analysis on the ADNI study. Theoretical properties of our estimators and their proofs, additional simulation and real data analysis results can be found in the Supplementary Material.

## 2. Data and problem description for ADNI.

The ADNI is a large-scale, multisite neuroimaging study that has collected clinical, imaging, genetic, and cognitive data at multiple time points from cognitively normal (CN) subjects, subjects with mild cognitive impairment (MCI), and AD patients (Mueller et al. (2005)). It supports the investigation and development of treatments that may slow or stop the progression of AD. The primary goal of ADNI is to test whether genetic, structural and functional neuroimaging, and clinical data can be integrated to assess the progression of MCI and early AD.

We constructed a dataset from the ADNI database (adni.loni.usc.edu), which consists of 606 subjects, including 113 AD patients, 316 patients with mild cognitive impairment (MCI), and 177 normal controls (NC). The dataset also includes demographic variables such as Age, Gender (0 = Male; 1 = Female), Handedness (0 = Right; 1 = Left), Retirement (0 = No; 1 = Yes), and Years of Education. The average age is 75.6 years with a standard deviation of 6.6 years, and the average years of education is 15.7 years with a standard deviation of 2.9 years. Among all the subjects, 361 are male, and 245 are female; 562 are right-handed, and 44 are left-handed; 497 are retired, and 109 are not.

We extracted GIC variables as follows. First, we extracted 13 cognitive variables at 12 months after the onset of ADNI for measuring the severity of cognitive impairment (Battista, Salvatore and Castiglioni (2017), Grassi et al. (2019)); see Table S1 in the Supplementary Material (Li et al. (2024)) for a summary of the abbreviations of these variables. Figure 2 presents the correlations between these scores. Among them, DIGITSCORE, RAVLT.learning, RAVLT.immediate, LDELTOTAL, and MMSE are positively correlated, with lower values indicating more severe cognitive impairment. In contrast, CDRSB, FAQ,

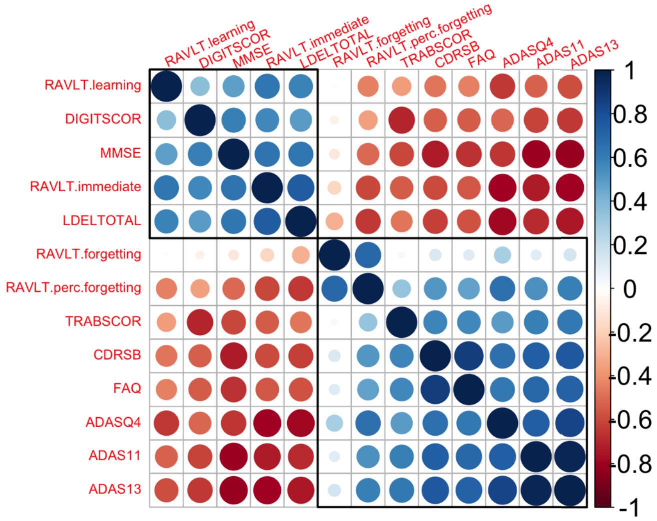


FIG. 2. *ADNI data analysis*: *correlations between the* 13 *cognitive variables*.

RAVLT.forgetting, RAVLT.perc.forgetting, TRABSCOR, ADAS11, ADAS13, and ADASQ4 are negatively correlated, with higher values indicating more severe cognitive impairment. Second, we used the image processing pipeline in (Li et al. (2007)) to calculate the hippocampal morphometry surface measure as a 100 × 150 matrix, with each element being a continuous variable representing the radial distance from the corresponding coordinate on the hippocampal surface to the medial core of the hippocampus. Such hippocampus surface measures may provide more subtle indexes, compared with volume differences in discriminating between patients with Alzheimer's and healthy control subjects. Third, we extracted ultrahigh dimensional genetic markers and other demographic covariates at baseline. We used the same preprocessing pipeline used in (Zhao et al. (2019, 2021, 2022)) to obtain 6,087,205 genotyped and imputed single-nucleotide polymorphisms (SNPs) on all of the 22 chromosomes.

The clinical spectrum of AD can be quite heterogeneous. Specifically, there are two main clinical syndromes: Amnestic AD, which is characterized by significant impairment of learning and recall, and nonamnestic AD, which involves impairment of language, visuospatial, or executive function (McKhann et al. (2011)). These scores measure different functions and may lack sensitivity at different stages of AD. For example, the detection of changes in ADAS11 and ADAS13 is limited by a substantial floor effect (Hobart et al. (2013)), whereas CDRSB lacks sensitivity to detect changes in the very early stages of AD (de Aquino (2021)).

Little is known about the genetic architecture of these cognitive scores and the genetic-imaging-clinical (GIC) pathway for AD. We are particularly interested in the following scientific questions:

- (Q1) How can we quantify the joint effect of genetic and imaging markers on the 13 cognitive scores?
- (Q2) How can we measure the shared and different heritability patterns of the 13 different cognitive scores with/without accounting for APOE4, which is considered to be the strongest risk factor gene for AD?
- (Q3) How do the estimates and heritability patterns of the 13 different cognitive scores differ with/without accounting for baseline diagnosis information?

We use model (1) to address (Q1)–(Q3) below.

## 3. Estimation and inference procedures for PFLR.

3.1. *Estimation algorithm.* In this subsection we develop an estimation method for model (1). First, we need to introduce some notation. Denote $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{X} = (X_1^T, \dots, X_n^T)^T$, $\mathbf{Z} = (Z_1, \dots, Z_n)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$. Denote the true value of $\xi$ and $\beta$ as $\xi^*$ and $\beta^*$, respectively. Let $S = \{1, 2, \dots, p\}$. For any $A$ and $B \subseteq S$ with length $|A|$ and $|B|$, denote $\beta_A = (\beta_i, i \in A) \in \mathbb{R}^{|A|}$. Denote $\beta|_A \in \mathbb{R}^p$ be a vector with its $i$th element $(\beta|_A)_i = \beta_i 1(i \in A)$, where $1(\cdot)$ is the indicator function. Let $\|\beta\|_{k,\infty}$ be the $k$th largest elements in absolute value. Denote $\|\cdot\|_0$ as the $\ell_0$ norm that calculates the number of nonzero elements of a vector. Let $\|\cdot\|_2$ be the $\ell_2$-norm such that $\|\beta\|_2^2 = \sum_{i=1}^p \beta_i^2$ and $\|\cdot\|_{L_2}$ be the $L_2$-norm such that $\|\xi\|_{L_2}^2 = \int_{\mathcal{T}} \xi^2(t)\,dt$. Thus, we assume throughout $E(X) = 0$, $E\{Z(t)\} = 0$ and $E(Y) = 0$, and, therefore, the intercept term can be ignored. In practice, we assume that the response $Y_i$ and the predictors $X_i$ and $Z_i(\cdot)$ are all mean centered such that $n^{-1}\sum_{i=1}^n Y_i = 0$, $n^{-1}\sum_{i=1}^n X_i = \mathbf{0}$, and $n^{-1}\sum_{i=1}^n Z_i(\cdot) = 0$.

We use the least-squares loss to estimate the functional and scalar coefficients. Due to the infinite-dimensional functional coefficient and high-dimensional scalar coefficients, regularizations are needed for estimating both $\xi(t)$ and $\beta$. Similar to Yuan and Cai (2010) and Cai and Yuan (2011), the functional coefficient $\xi^*$ is assumed to reside in a RKHS $\mathcal{H}(K)$ with a reproducing kernel $K$. The RKHS roughness penalty is imposed on the functional parameter, while the $\ell_0$ penalty is imposed on the scalar parameters, following a similar spirit of Huang et al. (2018). Therefore, we solve the following minimization problem:

$$\min_{\beta \in \mathbb{R}^p, \xi \in \mathcal{H}} (2n)^{-1} \sum_{i=1}^n \left[ Y_i - \left( X_i^T \beta + \int_{\mathcal{T}} Z_i(t)\xi(t)\mathrm{d}t \right) \right]^2 \quad \text{subject to } \|\beta\|_0 \le J,\ \|\xi\|_{\mathcal{H}}^2 \le \widetilde{J},$$

where $J > 0$ controls the sparsity level of $\beta$ and $\widetilde{J} > 0$ controls the smoothness level of $\xi$. Consider the Lagrangian form of the above minimization problem

$$(2) \qquad \min_{\beta \in \mathbb{R}^p, \xi \in \mathcal{H}} \left\{ (2n)^{-1} \sum_{i=1}^n \left[ Y_i - \left( X_i^T \beta + \int_{\mathcal{T}} Z_i(t)\xi(t)\mathrm{d}t \right) \right]^2 + \tau\|\beta\|_0 + 0.5\lambda\|\xi\|_{\mathcal{H}}^2 \right\},$$

where $\tau$ and $\lambda$ are the Lagrange multipliers. To solve the minimization problem (2), the following Representer theorem is very useful.

THEOREM 1. *For any $\beta \in \mathbb{R}^p$, there exists a parameter vector $\mathbf{c}(\beta)$ such that*

$$(3) \qquad \widehat{\xi}(\beta) = \sum_{i=1}^n c_i(\beta)(KZ_i),$$

*where $c_i(\beta)$ is the $i$th component of $\mathbf{c}(\beta)$ and $Kf := \int_{\mathcal{S}} K(\cdot, t) f(t)\mathrm{d}t$ for $f \in L_2(\mathcal{T})$.*

Define $\boldsymbol{\Sigma}_{ii'} = \iint_{\mathcal{S}\times\mathcal{S}} Z_i(s)K(s,t)Z_{i'}(t)\mathrm{d}s\mathrm{d}t$ as the $(i, i')$-th entry of $\boldsymbol{\Sigma}$. The objective function in (2) can be written in matrix form as

$$(4) \qquad (2n)^{-1}\|\mathbf{Y} - \mathbf{X}\beta - \boldsymbol{\Sigma}\mathbf{c}\|_2^2 + \tau\|\beta\|_0 + 0.5\lambda\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}.$$

Taking the first-order derivative of (4) with respect to $\mathbf{c}$ and setting it to zero give

$$(5) \qquad \mathbf{c} = (\boldsymbol{\Sigma} + n\lambda\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\beta).$$

Substituting (5) into (4), we obtain the following minimization problem:

$$(6) \qquad \min_{\beta \in \mathbb{R}^p} \{(2n)^{-1}(\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{P}_\lambda(\mathbf{Y} - \mathbf{X}\beta) + \tau\|\beta\|_0\},$$

where $\mathbf{P}_\lambda := n\lambda(\boldsymbol{\Sigma} + n\lambda\mathbf{I})^{-1}$. Once we find an approximate solution to (6), we can plug it back into (5), leading to an estimate of $\xi$.

We derive the optimality conditions for (6). If $\beta^o$ is a minimizer of (6), then we have

$$d^o = \mathbf{X}^T\mathbf{P}_\lambda(\mathbf{Y} - \mathbf{X}\beta^o)/n \quad \text{and} \quad \beta^o = H_\tau(\beta^o + d^o), \tag{7}$$

where $H_\tau(\cdot)$ is the elementwise hard thresholding operator with its $i$th entry, defined by $H_\tau(\beta)_i = 0$, if $|\beta_i| < \sqrt{2\tau}$ and $\beta_i$, if $|\beta_i| \geq \sqrt{2\tau}$. Conversely, if $\beta^o$ and $d^o$ satisfy (7), then $\beta^o$ is a local minimizer of (6).

Let $A^o = \{i : \beta_i \neq 0\}$ and $I^o = \{i : \beta_i = 0\}$. Denote $\beta^o_{A^o} = (\beta_i : i \in A^o) \in \mathbb{R}^{|A^o|}$ and, similarly, $\beta^o_{I^o}$, $d^o_{A^o}$, and $d^o_{I^o}$. Denote $\mathbf{X}_{A^o} = (\mathbf{X}_i : i \in A^o) \in \mathbb{R}^{n\times|A^o|}$ and, similarly, $\mathbf{X}_{I^o}$. By (7) we have $A^o = \{i : |\beta^o_i + d^o_i| \geq \sqrt{2\tau}\}$, $I^o = \{i : |\beta^o_i + d^o_i| < \sqrt{2\tau}\}$, and the system of equations

$$\beta^o_{I^o} = \mathbf{0}, \qquad d^o_{A^o} = \mathbf{0}, \beta^o_{A^o} = (\mathbf{X}^T_{A^o}\mathbf{P}_\lambda\mathbf{X}_{A^o})^{-1}\mathbf{X}^T_{A^o}\mathbf{P}_\lambda\mathbf{Y}, \qquad d^o_{I^o} = \mathbf{X}^T_{I^o}\mathbf{P}_\lambda(\mathbf{Y} - \mathbf{X}_{A^o}\beta^o_{A^o})/n.$$

If we want $\beta^o$ to have exactly $J$ nonzero elements, then we can set $\sqrt{2\tau}$ equal to the $J$th largest element of the sequence $\{|\beta^o_i + d^o_i| : i = 1, \dots, p\}$.

To solve the above system of equations and obtain both the functional and scalar estimates, for a given sparsity level $J$, we modify the support detection and root finding algorithm (Huang et al. (2018)) to accommodate the functional variable. We use the generalized cross-validation (GCV) to select the tuning parameter $\lambda$ by following the same reasoning in Yuan and Cai (2010). To select an appropriate sparsity level $J$, we use the high-dimensional Bayesian information criterion (HBIC) (Wang, Kim and Li (2013)) as follows:

$$\mathrm{HBIC}_J = \log\left(n^{-1}\left\|\mathbf{Y} - \mathbf{X}\widehat{\beta} - \int_{\mathcal{T}} \mathbf{Z}(t)\widehat{\xi}(t)\,dt\right\|^2\right) + J\log\log(n)\log(p)/n.$$

The sparsity level $J$ is set to minimize $\mathrm{HBIC}_J$.

We summarize the above algorithm in Algorithm 1. The proposed estimation algorithm can be easily extended to allow for certain sets of covariates not to be penalized by including these covariates in the active set at each iteration.

**Algorithm 1** Functional support detection and root finding (FSDAR)

```
Input: An initial $\beta^0$ and the sparsity level $J$; set $k = 0$.
for a given $\lambda$, calculate $d^0 = \mathbf{X}^T\mathbf{P}_\lambda(\mathbf{Y} - \mathbf{X}\beta^0)/n$ and $\mathbf{P}_\lambda = n\lambda(\boldsymbol{\Sigma} + n\lambda\mathbf{I})^{-1}$;
for $k = 0, 1, 2, \dots$ do
    $A^k = \{i : |\beta^k_i + d^k_i| \geq \|\beta^k + d^k\|_{J,\infty}\}$, $I^k = (A^k)^c$;
    $\beta^{k+1}_{A^k} = (\mathbf{X}^T_{A^k}\mathbf{P}_\lambda\mathbf{X}_{A^k})^{-1}\mathbf{X}^T_{A^k}\mathbf{P}_\lambda\mathbf{Y}$, $\beta^{k+1}_{I^k} = \mathbf{0}$;
    $d^{k+1}_{A^k} = \mathbf{0}$, $d^{k+1}_{I^k} = \mathbf{X}^T_{I^k}\mathbf{P}_{\lambda^k}(\mathbf{Y} - \mathbf{X}_{A^k}\beta^{k+1}_{A^k})/n$;
    if $A^{k+1} = A^k$ then
        Stop and denote $\widehat{\beta} = (\widehat{\beta}^T_{A^k}, \widehat{\beta}^T_{I^k})^T$.
    else
        $k = k + 1$;
    end if
end for
for a list of candidate $\lambda$s, repeat steps 1-11 and select $\lambda$ that gives the best GCV = $n\|\mathbf{P}_\lambda(\mathbf{Y} - \mathbf{X}_{A^k}\beta^k_{A^k})\|_2^2/[\mathrm{tr}(\mathbf{P}_\lambda)]^2$
Output: $\widehat{\beta}$, $\widehat{\mathbf{c}} = (\boldsymbol{\Sigma} + n\lambda\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\beta})$, and $\widehat{\xi} = \sum_{i=1}^n \widehat{\mathbf{c}}_i(KZ_i)$.
```

3.2. *Computational complexity.* We analyze the computational complexity of Algorithm 1 as follows. Let $m$ denote the number of points of the functional variable. The calculation of the matrix $\mathbf{\Sigma}$ has a complexity of $O(nm^2 + n^2m)$ and can be precomputed and stored. The computational complexity of calculating the matrix $\mathbf{P}_\lambda$ is $O(nm^2 + n^2m + n^3)$, while the complexity of calculating $d^0$ is $O(n^2p)$. Therefore, the complexity of step 1 is $O(n^2p)$, assuming that $p$ is much larger than the sample size. Step 3 requires $O(p)$ flops, step 4 and step 5 both have a complexity of $O(n^2J)$, and checking the stopping condition in step 6 takes $O(p)$. Consequently, for a given $\lambda$ and sparsity level $J$, the overall cost per iteration of Algorithm 1 is $O(n^2p)$. According to Corollary 2 in the Supplementary Material (Li et al. (2024)), no more than $O(\log(R))$ iterations are needed to obtain a good solution, where $R = \max\{|\beta_i^*|, i \in A^*\}/\min\{|\beta_i^*|, i \in A^*\}$. Therefore, the overall cost of Algorithm 1 is $O(n^2p\log(R))$. If we select the tuning parameters $\lambda$, $J$ using a grid search method, the computational complexity increases by a factor corresponding to the number of tuning parameter sets.

3.3. *Inference procedure.* In order to provide theoretical guarantees for all estimators and test statistics, we establish their theoretical properties in the Supplementary Material (Li et al. (2024)). These properties encompass the nonasymptotic error bounds of $\widehat{\beta}_k$ at each iteration, the general nonasymptotic error bound of $\widehat{\xi}(t)$, the asymptotic normality of the estimates for the nonzero scalar coefficients, and the null limit distribution of $T_\xi$.

As an example, we consider testing the nullity of the functional covariate $\xi(t)$. In our real data analysis, we aim to examine whether the left and right hippocampi have significant effects on cognitive decline. To do so, we propose the following test:

$$H_0 : \xi_0(t) = 0 \quad \text{for any } t, \quad \text{vs.} \quad H_1 : \xi_0(t) \neq 0, \quad \text{for some } t. \tag{8}$$

After variable selection we adapt the test statistic in Li and Zhu (2020) to test (8). Specifically, we define

$$T_\xi = 2n\big[\ell_{n\lambda}\big(\hat{\beta}_{H_0}, \xi_0(t)\big) - \ell_{n\lambda}\big(\hat{\beta}, \hat{\xi}(t)\big)\big],$$

where $\ell_{n\lambda}$ is the loss function and $\hat{\beta}_{H_0}$ is the estimator under the null hypothesis. Following Li and Zhu (2020), we can demonstrate that $T_\xi$ converges to a normal distribution, while being approximated a chi-square distribution, as shown in Corollary 4 of the Supplementary Material (Li et al. (2024)).

**4. Simulation studies.** We investigate the finite sample performance of the proposed estimation method in two cases: one-dimensional $\xi(s)$, which is discussed in this section, and two-dimensional $\xi(s)$, which is presented in the Supplementary Material (Li et al. (2024)). We employ Algorithm 1 to estimate the unknown coefficients. The initial value is set to zero, and we choose $J \in 1, 2, \dots, 50$ and $\lambda$ from 50 evenly spaced points on the interval $[1e - 5, 0.1]$.

EXAMPLE 4.1. The following is designed to evaluate the estimation and prediction performances for one-dimensional $\xi(t)$ in the interval $\mathcal{T} = [0, 1]$. The functional predictor $Z(t)$ is given by: $Z(t) = \sum_{k=1}^{50} U_k\phi_k(t)$ for $t \in [0, 1]$, where $\phi_{2l-1}(t) = \sqrt{2}\cos((2l - 1)\pi t)$ and $\phi_{2l}(t) = \sqrt{2}\sin((2l - 1)\pi t)$, $\text{l} = 1, \dots, 25$, and $\{U_k\}$ are independently sampled from the normal distribution $N(0, 16|k - C_0| + 1)$ with $C_0 \in \{1, 3\}$. For the coefficient function, we set $\xi(t) = \sum_{k=1}^{50} 4(-1)^{k+1}k^{-2}\phi_k(t)$. When $C_0 = 1$, the functional coefficient can be efficiently represented in terms of the leading functional principal components. When $C_0 = 3$, the representative basis functions for $Z(t)$ and $\xi(t)$ are disordered such that the leading eigenfunctions $\phi_k(t)$ of the covariance kernel of $Z(t)$ are around $k = 3$.

Following Kong et al. (2016), we allow a moderate correlation between $Z(t)$ and the scalar covariates $X = (X_1, \dots, X_p)^T$ by introducing a correlation structure between $\{U_1, U_2, U_3, U_4\}$ and $X = (X_1, \dots, X_p)^T$ as $\mathrm{corr}(U_k, X_l) = \rho_1^{|k-l|+1}$ for $k = 1, \dots, 4$ and $l = 1, \dots, p$ with $\rho_1 \in \{0.2, 0.4\}$. The scalar covariates $X = (X_1, \dots, X_p)^T$ are jointly normal with zero mean, unit variance, and $\mathrm{AR}(\rho_2)$ with $\rho_2 \in \{0.3, 0.5, 0.7\}$.

For each subject $i$, we observe $Z_i(t_{ij})$ at 100 equally spaced points. The errors $\epsilon_i$s are generated from the standard normal distribution. The sample size is chosen to be $n = 200$. We consider $\beta$ with two different values of dimensionality: $p = 150$, which is *smaller* than the training sample size, and $p = 1500$, which is *larger* than the training sample size. Specifically, the underlying true $\beta$ is set to be $\beta = (3, 1.5, 1, 2.5, 2, \underbrace{0, \dots, 0}_{p-5})^T$. In addition to the proposed method, the method based on FPCA proposed by Kong et al. (2016) is also considered for comparison. The number of the functional components and the penalty tuning parameter are selected by minimizing the value of HBIC. The initial $\beta^0$ is set to be 0. We have tested several different choices, including $\beta^0 = (1, \dots, 1)^\top$, $\beta^0 = (10, \dots, 10)^\top$, $\beta^0 = (100, \dots, 100)^\top$, and the marginal correlation between the covariates and the response. The results are the same, which are not sensitive to the selection of the initial $\beta^0$.

All simulation results are based on 200 replications by using R (version 3.6.0) on a Linux server (equipped with Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40 GHz, 125 GB RAM). We evaluate the estimation accuracy of $\widehat{\beta}$ using the mean squared error $\mathrm{MSE}_\beta = \|\widehat{\beta} - \beta\|_2^2$ and that of $\xi$ by using the mean integrated squared error $\mathrm{MSE}_\xi = \|\widehat{\xi} - \xi\|_{L_2}^2$ as well as the relative MSE of $\widehat{\xi}$ such that $\mathrm{RMSE}_\xi = \|\widehat{\xi} - \xi\|_{L_2}^2 / \|\xi\|_{L_2}^2$. We also calculate the number of false zero scalar predictors (FZ), the number of false nonzero scalar predictors (FN), and the prediction mean squared error (PMSE) based on 200 new test samples. We further calculate the compatation time (in seconds).

Table 1 presents the variable selection accuracy, estimation accuracy, and prediction results for the moderate number of scalar variables with $n = 200$ and $p = 150$. Our method outperforms the FPCA-based method in Kong et al. (2016) in almost all scenarios. Specifically, the selection of scalar predictors for our method is more accurate and stable than the competing method, with fewer false nonzero scalars and zero false zero scalars. For our method, the number of false zero scalars and false nonzero scalars do not differ too much across different correlations among the scalar variables. However, FZ and FN of the competing method (Kong et al. (2016)) increase as the correlation among the scalar variables becomes larger. This indicates that more zero scalar variables would be selected in PFLR, whereas more nonzero scalar variables would be excluded from PFLR. When the representative basis functions for $Z(t)$ and $\xi(t)$ are not exactly matched, our method still yields more stable estimates than the competing method (Kong et al. (2016)). Furthermore, MSEs and PMSEs for our method are smaller than those for the competing method (Kong et al. (2016)) in all scenarios.

Table 2 reports additional simulation results corresponding to $n = 200$ and $p = 1500$. The proposed method outperforms the competing method (Kong et al. (2016)) in terms of FNs, FZs, MSEs, and PMSEs. For instance, it is noteworthy that the number of false zero scalars for the competing method (Kong et al. (2016)) increases as $\rho_2$ increases.

Figure 3 shows the solution path of $\widehat{\beta}$ for different $\rho_2$ values corresponding to $(p, C_0, \rho_1) = (150, 1, 0.2)$. The solution path displays how $\widehat{\beta}$ evolves either as the sparsity level of the proposed method increases or as the penalty tuning parameter decreases. Specifically, the upper five lines show how $\widehat{\beta_1}, \dots, \widehat{\beta_5}$ change with the sparsity level and the penalty tuning parameter. For our method, when $\rho_2$ is small (e.g., $\rho_2 = 0.3$), the five variables gradually enter the model and the more significant a variable is, the earlier it is selected. However, when $\rho_2$ increases to a larger value, a less important variable may enter the model earlier than a

TABLE 1

*Simulation results of Monte Carlo averages and empirical standard errors in parentheses for $n = 200,\ p = 150$*

| Center | $\rho_1$ | $\rho_2$ | | FZ | FN | $\text{MSE}_\beta$ | $\text{MSE}_\xi$ | $\text{RMSE}_\xi$ | PMSE | Time(s) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.2 | 0.3 | Proposed | 0.000(0.000) | 0.670(0.737) | 0.067(0.046) | 0.035(0.023) | 0.002(0.001) | 1.085(0.121) | 2.219(2.458) |
| | | | FPCA | 0.005(0.071) | 1.115(2.875) | 0.152(0.307) | 0.103(0.051) | 0.006(0.003) | 1.269(0.273) | 0.590(0.098) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.685(0.767) | 0.082(0.056) | 0.036(0.026) | 0.002(0.001) | 1.087(0.122) | 2.714(2.987) |
| | | | FPCA | 0.015(0.122) | 1.390(2.890) | 0.224(0.346) | 0.103(0.051) | 0.006(0.003) | 1.271(0.228) | 0.694(0.122) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.530(0.679) | 0.107(0.072) | 0.036(0.024) | 0.002(0.001) | 1.076(0.123) | 2.737(2.744) |
| | | | FPCA | 0.255(0.437) | 1.350(3.015) | 0.779(1.874) | 0.110(0.055) | 0.006(0.003) | 1.377(1.083) | 0.685(0.125) |
| | 0.4 | 0.3 | Proposed | 0.000(0.000) | 0.695(0.731) | 0.072(0.046) | 0.039(0.029) | 0.002(0.002) | 1.081(0.114) | 2.572(2.537) |
| | | | FPCA | 0.015(0.122) | 1.150(2.594) | 0.183(0.367) | 0.115(0.065) | 0.007(0.004) | 1.284(0.349) | 0.701(0.123) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.635(0.703) | 0.079(0.056) | 0.038(0.032) | 0.002(0.002) | 1.083(0.115) | 2.981(2.930) |
| | | | FPCA | 0.035(0.184) | 1.155(2.448) | 0.304(0.523) | 0.113(0.062) | 0.006(0.004) | 1.302(0.306) | 0.678(0.120) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.510(0.680) | 0.111(0.079) | 0.038(0.024) | 0.002(0.001) | 1.078(0.118) | 2.423(2.518) |
| | | | FPCA | 0.300(0.470) | 1.020(2.143) | 0.658(1.016) | 0.122(0.066) | 0.007(0.004) | 1.310(0.315) | 0.704(0.125) |
| 3 | 0.2 | 0.3 | Proposed | 0.000(0.000) | 0.655(0.706) | 0.067(0.046) | 0.026(0.016) | 0.001(0.001) | 1.088(0.118) | 2.527(2.441) |
| | | | FPCA | 0.000(0.000) | 0.125(0.448) | 0.052(0.076) | 0.193(0.103) | 0.011(0.006) | 1.369(0.185) | 0.712(0.134) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.665(0.752) | 0.081(0.055) | 0.026(0.017) | 0.001(0.001) | 1.086(0.123) | 2.338(2.231) |
| | | | FPCA | 0.000(0.000) | 0.250(0.714) | 0.065(0.067) | 0.191(0.102) | 0.011(0.006) | 1.365(0.178) | 0.702(0.140) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.520(0.687) | 0.105(0.073) | 0.025(0.015) | 0.001(0.001) | 1.080(0.124) | 2.765(2.791) |
| | | | FPCA | 0.050(0.218) | 0.340(0.805) | 0.196(0.377) | 0.187(0.095) | 0.011(0.005) | 1.369(0.190) | 0.683(0.122) |
| | 0.4 | 0.3 | Proposed | 0.000(0.000) | 0.650(0.728) | 0.068(0.045) | 0.028(0.019) | 0.002(0.001) | 1.081(0.118) | 2.396(2.307) |
| | | | FPCA | 0.000(0.000) | 0.090(0.335) | 0.053(0.049) | 0.190(0.106) | 0.011(0.006) | 1.361(0.174) | 0.667(0.118) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.555(0.692) | 0.072(0.055) | 0.027(0.017) | 0.002(0.001) | 1.080(0.113) | 2.637(2.744) |
| | | | FPCA | 0.000(0.000) | 0.195(0.573) | 0.073(0.079) | 0.197(0.099) | 0.011(0.006) | 1.368(0.163) | 0.668(0.124) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.530(0.715) | 0.107(0.082) | 0.027(0.015) | 0.002(0.001) | 1.081(0.118) | 2.795(2.912) |
| | | | FPCA | 0.110(0.314) | 0.225(0.553) | 0.245(0.456) | 0.201(0.125) | 0.011(0.007) | 1.389(0.209) | 0.664(0.120) |

TABLE 2

*Simulation results of Monte Carlo averages and empirical standard errors in parentheses for $n = 200$, $p = 1500$*

| Center | $\rho_1$ | $\rho_2$ | | FZ | FN | $\text{MSE}_\beta$ | $\text{MSE}_\xi$ | $\text{RMSE}_\xi$ | PMSE | Time(s) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.2 | 0.3 | Proposed | 0.000(0.000) | 0.820(0.813) | 0.098(0.068) | 0.094(0.011) | 0.005(0.001) | 1.191(0.135) | 35.566(33.777) |
| | | | FPCA | 0.005(0.071) | 4.655(9.863) | 0.224(0.763) | 0.107(0.078) | 0.006(0.004) | 1.344(0.812) | 1.079(0.233) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.855(0.811) | 0.121(0.075) | 0.094(0.012) | 0.005(0.001) | 1.207(0.143) | 38.249(34.465) |
| | | | FPCA | 0.035(0.184) | 5.190(9.596) | 0.429(1.111) | 0.111(0.057) | 0.006(0.003) | 1.418(0.964) | 1.090(0.250) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.665(0.778) | 0.150(0.098) | 0.092(0.010) | 0.005(0.001) | 1.178(0.137) | 38.350(34.027) |
| | | | FPCA | 0.410(0.493) | 5.480(9.252) | 1.031(1.295) | 0.115(0.059) | 0.007(0.003) | 1.465(0.760) | 1.094(0.219) |
| | 0.4 | 0.3 | Proposed | 0.000(0.000) | 0.715(0.766) | 0.108(0.065) | 0.100(0.013) | 0.006(0.001) | 1.183(0.138) | 39.534(34.614) |
| | | | FPCA | 0.010(0.100) | 4.560(9.637) | 0.233(0.616) | 0.118(0.067) | 0.007(0.004) | 1.346(0.620) | 1.047(0.247) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.675(0.736) | 0.121(0.075) | 0.097(0.012) | 0.005(0.001) | 1.174(0.133) | 37.203(32.144) |
| | | | FPCA | 0.065(0.247) | 5.205(8.870) | 0.501(1.111) | 0.125(0.063) | 0.007(0.004) | 1.425(0.783) | 1.096(0.236) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.685(0.767) | 0.179(0.113) | 0.095(0.011) | 0.005(0.001) | 1.173(0.143) | 35.889(31.236) |
| | | | FPCA | 0.455(0.509) | 4.795(8.094) | 1.089(1.434) | 0.139(0.086) | 0.008(0.005) | 1.440(0.562) | 1.090(0.223) |
| 3 | 0.2 | 0.3 | Proposed | 0.000(0.000) | 0.205(0.473) | 0.081(0.068) | 0.073(0.017) | 0.004(0.001) | 1.467(0.185) | 36.373(33.003) |
| | | | FPCA | 0.000(0.000) | 0.395(0.961) | 0.057(0.059) | 0.217(0.135) | 0.012(0.008) | 1.391(0.178) | 1.045(0.235) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.240(0.494) | 0.108(0.076) | 0.072(0.016) | 0.004(0.001) | 1.471(0.188) | 39.177(32.842) |
| | | | FPCA | 0.005(0.071) | 0.940(1.565) | 0.103(0.201) | 0.224(0.134) | 0.013(0.008) | 1.413(0.196) | 1.078(0.270) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.170(0.427) | 0.169(0.132) | 0.071(0.018) | 0.004(0.001) | 1.459(0.181) | 37.563(33.131) |
| | | | FPCA | 0.130(0.337) | 1.910(2.755) | 0.274(0.484) | 0.219(0.127) | 0.012(0.007) | 1.416(0.216) | 1.079(0.203) |
| | 0.4 | 0.3 | Proposed | 0.000(0.000) | 0.255(0.549) | 0.146(0.081) | 0.088(0.021) | 0.005(0.001) | 1.458(0.183) | 34.386(31.798) |
| | | | FPCA | 0.000(0.000) | 0.485(1.080) | 0.072(0.072) | 0.222(0.139) | 0.013(0.008) | 1.402(0.185) | 1.078(0.240) |
| | | 0.5 | Proposed | 0.000(0.000) | 0.235(0.540) | 0.164(0.100) | 0.082(0.018) | 0.005(0.001) | 1.452(0.182) | 38.020(32.011) |
| | | | FPCA | 0.010(0.100) | 0.905(1.568) | 0.133(0.229) | 0.227(0.137) | 0.013(0.008) | 1.420(0.204) | 1.061(0.228) |
| | | 0.7 | Proposed | 0.000(0.000) | 0.205(0.463) | 0.277(0.166) | 0.078(0.018) | 0.004(0.001) | 1.448(0.182) | 39.756(33.367) |
| | | | FPCA | 0.195(0.397) | 1.685(2.270) | 0.388(0.600) | 0.236(0.137) | 0.013(0.008) | 1.434(0.221) | 1.083(0.223) |

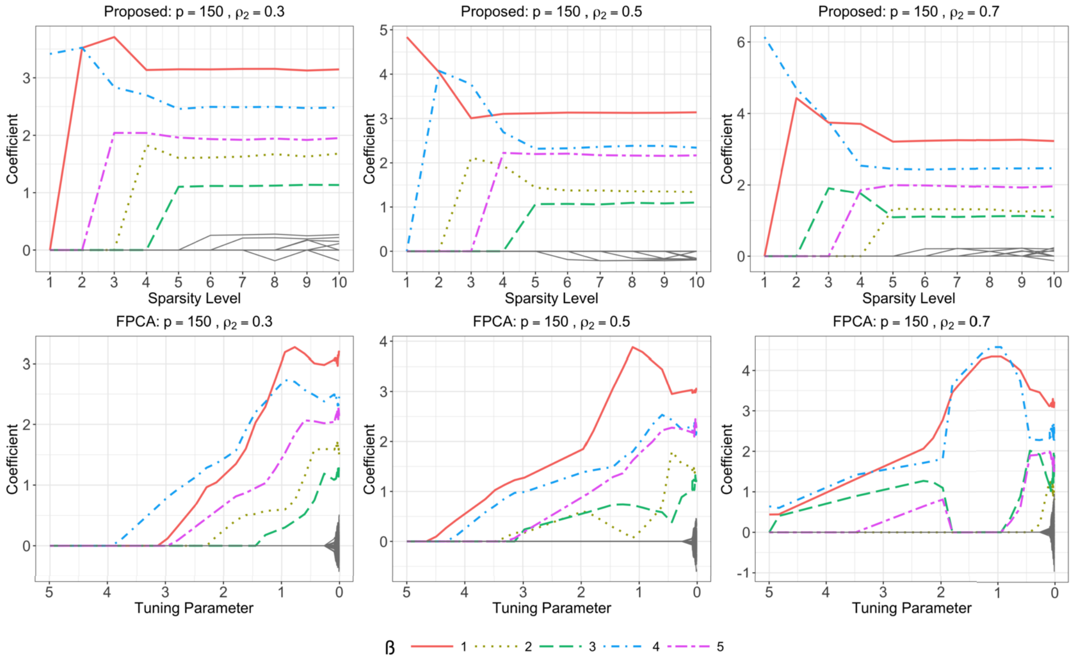


FIG. 3. *Solution paths of* $\widehat{\beta}$ *for different settings of for different* $\rho_2$ *corresponding to* $(p, C_0, \rho_1) = (150, 1, 0.2)$. *The upper five lines correspond to the solution paths of the nonzero elements of* $\beta$, *whereas the others correspond to zero elements.*

more important one. The evolution of significant $\widehat{\beta}$s does not change very much when the sparsity level is greater than the true sparsity level of 5. For the competing method (Kong et al. (2016)), we observed similar entering orders of the variables. The values of significant $\widehat{\beta}$s also do not differ very much when the tuning parameter is smaller than some value when $\rho_2$ is small. However, when the correlation among the scalar variables increases, the evolution of significant $\widehat{\beta}$s shows different patterns, such that nonzero scalar variables may be excluded from PFLR as the tuning parameter of the SCAD penalty decreases. These results may indicate that our method is more stable and accurate than the competing method.

EXAMPLE 4.2. In this example we evaluate the Type I and II error rates of the proposed test statistic. Data settings are the same as in Example 4.1, except that $\xi(t) = \sum_{k=1}^{50} B(-1)^{k+1}k^{-2}\phi_k(t)$ with $B \in \{0, 0.01, 0.03, 0.05, 0.07, 0.1\}$, which controls the signal strength. When $B = 0$, we obtain the sizes. Because the testing results have similar patterns for different values of $(\rho_1, \rho_2)$, we report the sizes and powers when $(\rho_1, \rho_2) = (0.2, 0.5)$ for the sake of a concise presentation. We choose $n \in \{200, 400\}$, $p \in \{150, 1500\}$, and the significance level to be 5%.

For the null hypothesis, $H_0 : \xi(t) = 0$; Table 3 summarizes the sizes and powers of the proposed test based on 1000 simulation runs. It reveals that the empirical sizes are reasonably controlled around the nominal level, and the empirical power increases with the sample size $n$ as well as the signal strength.

**5. ADNI data analysis.** To address research questions (Q1)–(Q3), we developed three PFLR models named Models 1–3 to analyze the ADNI dataset as follows:

- Model 1, described in Section 5.1, considers each of the 13 cognitive scores at Month12 in Table S1 for $Y_i$, utilizes the hippocampus morphometry surface $Z_i(\cdot, \cdot)$ and includes the allele codes of screened SNPs, the set of demographic covariates at baseline detailed

TABLE 3
*Testing results of Monte Carlo averages with $\rho_1 = 0.2$, $\rho_2 = 0.5$*

| $n$ | $p$ | $B = 0$ | $B = 0.01$ | $B = 0.03$ | $B = 0.05$ | $B = 0.07$ | $B = 0.1$ |
|---|---|---|---|---|---|---|---|
| 200 | 150 | 0.057 | 0.111 | 0.112 | 0.817 | 0.969 | 1 |
| | 1500 | 0.053 | 0.117 | 0.117 | 0.784 | 0.955 | 0.999 |
| 400 | 150 | 0.055 | 0.137 | 0.137 | 0.982 | 0.999 | 1 |
| | 1500 | 0.057 | 0.115 | 0.115 | 0.978 | 0.999 | 1 |

in Section 2, and the top five principal components (PCs) of the whole genome data for correcting population stratification (Price et al. (2006)) as predictors $X_i$.

- Model 2, described in Section 5.2, is almost the same as Model 1 but excludes the SNPs in the 19q13.32 region from the candidate SNPs and includes the number of APOE4 gene copies as one of the controlling covariates.
- Model 3, described in Section 5.3, is almost the same as Model 2 but further includes the baseline diagnosis status as one of the controlling covariates.

These models help in analyzing ADNI to address the research questions effectively.

5.1. *GIC pathways.* We fit Model 1 to ADNI for addressing (Q1) as follows. Since the number of SNPs is significantly larger than the sample size, we first apply the sure independence screening approach (Fan and Lv (2008)) to reduce the number of candidate SNPs, controlling demographic variables and the top five PCs. We sort SNPs in decreasing order of their absolute correlations with each cognitive score and keep the top 1000 for each score. Combining the top 1000 SNPs of each score across all 13 scores results in 10,546 different SNPs in $X_i$. For easy comparison we standardize all SNPs, cognitive scores, and demographic variables, including Age and Years of education. As both left and right hippocampi have 2D radial distance measures and are asymmetric (Pedraza, Bowers and Gilmore (2004)), we apply our method to the left and right hippocampi separately. In the estimation procedure, controlling variables are not penalized and are always in the active set. The initial value is set to zero. We choose the sparsity level $J$ and the smoothness parameter $J$ by a grid search method with $J \in 1, 2, \dots, 100$ and $\lambda$ from 50 evenly spaced points on $[1e - 5, 0.1]$.

Figures 4 and 5 present the estimates of the left and right hippocampal surfaces for all 13 cognitive scores in Model 1. Figure 4 shows the estimates $\widehat{\xi}$, with values ranging from 0.071 to 0.63, corresponding to DIGITSCOR, LDELTOTAL, MMSE, RAVLT.immediate, and RAVLT.learning. Figure 5 displays the estimates $\widehat{\xi}$, with values ranging from $-0.62$ to $-0.018$, corresponding to ADAS11, ADAS13, ADASQ4, CDRSB, FAQ, RAVLT.forgetting, RAVLT.perc.forgetting, and TRABSCOR. Examining Figures 4 and 5 reveals the heterogeneous effects of the hippocampus on all 13 cognitive scores. A bilateral and asymmetric hippocampal effect on cognitive function is observed. Among the six hippocampal subfields in Figure 4, CA1, presubiculum, and subiculum show high sensitivity (Frisoni et al. (2008), de Flores, Joie and Chételat (2015)). AD-related atrophy initially appears focal in CA1 before spreading to other subfields.

Hereafter we focus on the results obtained from the left hippocampal surface data. Table 4 presents the estimates of selected covariates and their corresponding $p$-values for the 13 cognitive scores in Model 1. The $p$-values are calculated using the limit distribution in Theorem 3 of the Supplementary Material (Li et al. (2024)), assuming that the true value of the coefficient is zero. Under this assumption each estimated coefficient converges to a normal distribution, and we use this distribution to determine the $p$-value. Consistent with

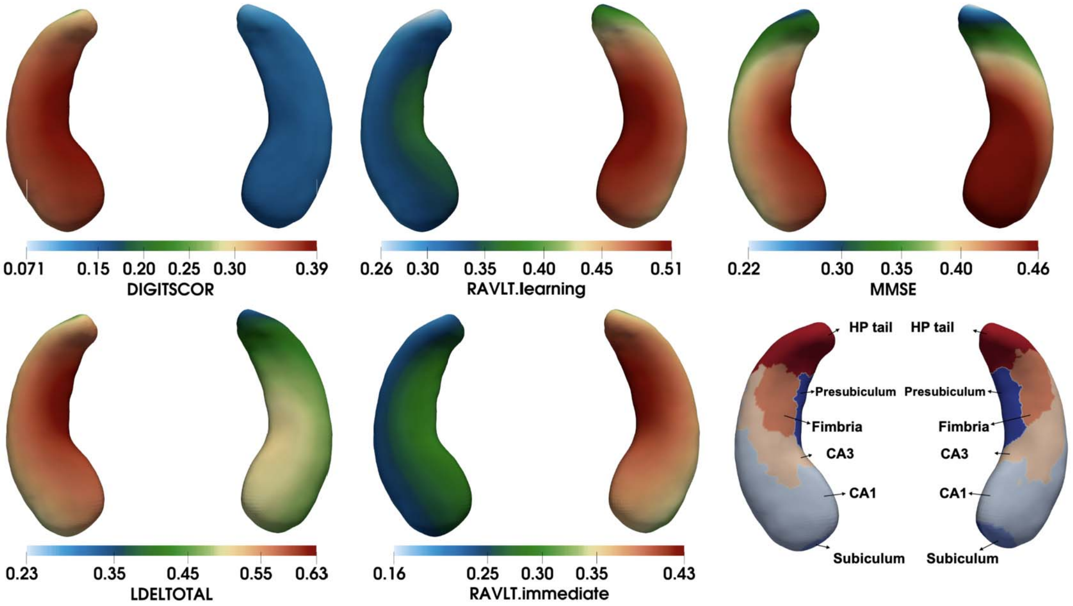


FIG. 4. *ADNI data analysis results*: *Estimates of the left and right hippocampus surfaces for DIGITSCOR*, *LDELTOTAL*, *MMSE*, *RAVLT.immediate*, *and RAVLT.learning for Model* 1 *and the hippocampal subfields* (*from left to right*, *and from top to bottom*).

previous research (Vina and Lloret (2010), Guerreiro and Bras (2015)), we find that age and education have a significant impact on most of the cognitive scores. Age generally has a negative effect on cognitive function, while education has positive effects depending on cognitive function. Retirement is significant for six scores, suggesting that retired individuals are at an increased risk of cognitive impairment. Gender is significant for five scores, and handedness is significant for three scores.

Figures 6(a) and 6(b) display the ideogram and Manhattan plots for significant SNPs related to all 13 cognitive scores in Model 1, respectively. Upon examining Figure 6(a), heterogeneous genetic effects are observed across all scores. However, several well-known SNPs

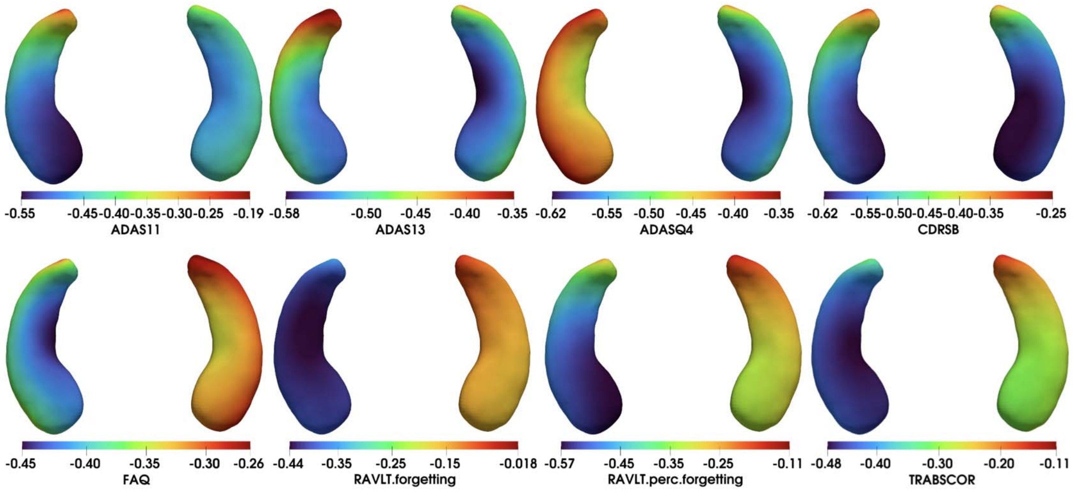


FIG. 5. *ADNI data analysis results*: *Estimates of the the left and right hippocampus surfaces for ADAS*11, *ADAS*13, *ADASQ*4, *CDRSB*, *FAQ*, *RAVLT.forgetting*, *RAVLT.perc.forgetting*, *and TRABSCOR for Model* 1 (*from left to right*, *and from top to bottom*).

TABLE 4

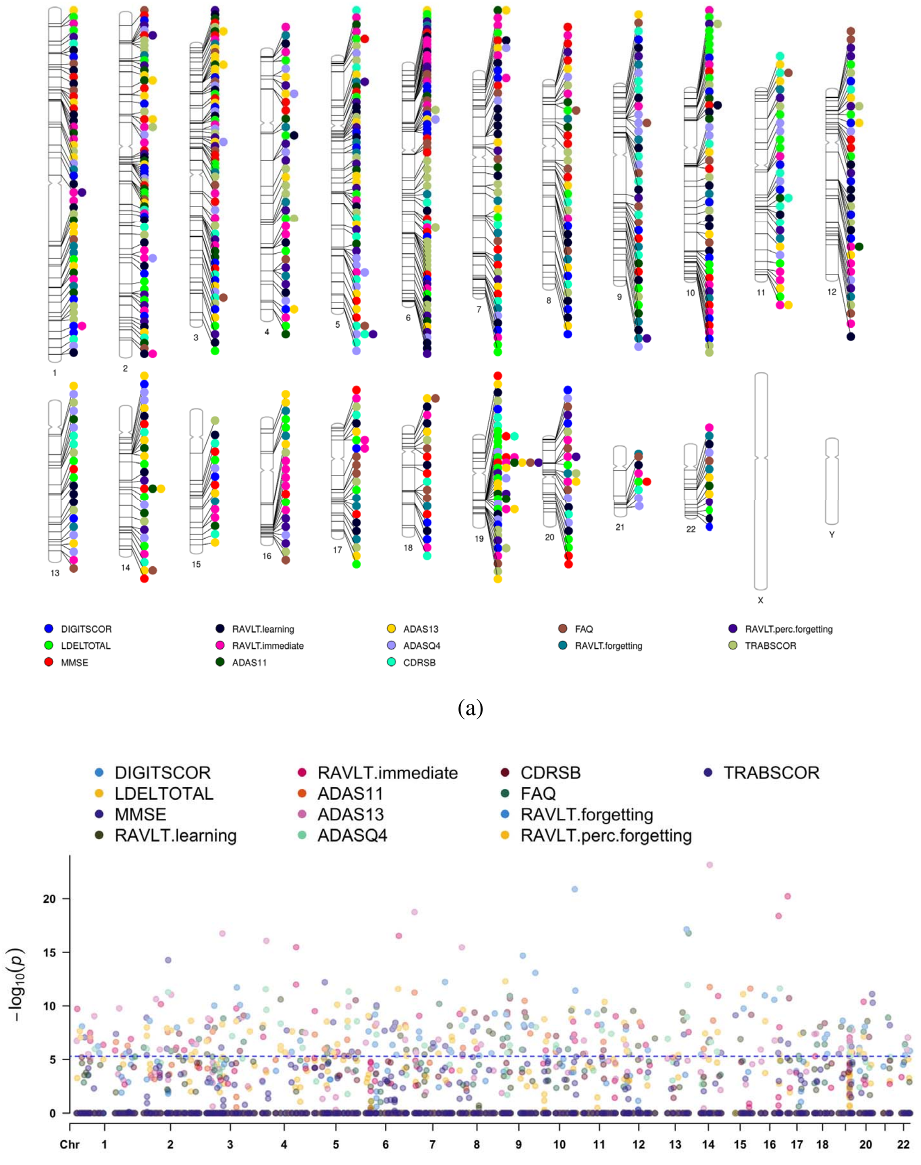
*ADNI data analysis results*: *estimates with their corresponding raw p-values in parentheses of some selected covariates for the* 13 *cognitive scores*

| Score | Model | Gender | Handedness | Education | Retirement | Age | APOE4 | MCI | AD |
|---|---|---|---|---|---|---|---|---|---|
| ADAS11 | Model 1 | 0.098(0.033) | 0.110(0.204) | −0.078(0.001) | 0.137(0.021) | 0.051(0.028) | – | – | – |
| | Model 2 | 0.129(0.000) | 0.145(0.012) | −0.091(0.000) | 0.070(0.088) | 0.074(0.000) | 0.252(0.000) | – | – |
| | Model 3 | 0.136(0.000) | 0.051(0.314) | −0.017(0.226) | −0.019(0.585) | 0.062(0.000) | 0.066(0.000) | 0.823(0.000) | 1.855(0.000) |
| ADAS13 | Model 1 | −0.034(0.333) | 0.123(0.057) | −0.124(0.000) | 0.007(0.872) | 0.050(0.005) | – | – | – |
| | Model 2 | 0.099(0.005) | 0.084(0.207) | −0.176(0.000) | 0.158(0.000) | 0.040(0.023) | 0.291(0.000) | – | – |
| | Model 3 | 0.164(0.000) | 0.004(0.929) | 0.028(0.021) | 0.010(0.752) | 0.043(0.001) | 0.077(0.000) | 0.994(0.000) | 1.967(0.000) |
| ADASQ4 | Model 1 | −0.058(0.100) | 0.161(0.016) | −0.138(0.000) | 0.042(0.348) | −0.035(0.051) | – | – | – |
| | Model 2 | −0.026(0.452) | 0.135(0.037) | −0.169(0.000) | 0.100(0.026) | 0.007(0.698) | 0.308(0.000) | – | – |
| | Model 3 | 0.007(0.757) | 0.141(0.002) | −0.091(0.000) | −0.001(0.969) | 0.006(0.612) | 0.173(0.000) | 1.118(0.000) | 1.746(0.000) |
| CDRSB | Model 1 | 0.019(0.660) | 0.097(0.220) | −0.091(0.000) | 0.132(0.017) | 0.047(0.027) | – | – | – |
| | Model 2 | 0.085(0.011) | −0.0.089(0.172) | −0.081(0.000) | 0.265(0.004) | 0.053(0.002) | 0.215(0.000) | – | – |
| | Model 3 | 0.107(0.000) | −0.029(0.494) | 0.016(0.154) | 0.107(0.000) | 0.046(0.000) | 0.071(0.000) | 0.819(0.000) | 2.017(0.000) |
| FAQ | Model 1 | 0.050(0.232) | 0.052(0.519) | −0.079(0.000) | 0.278(0.000) | 0.045(0.035) | – | – | – |
| | Model 2 | 0.145(0.004) | 0.020(0.830) | −0.067(0.007) | 0.273(0.000) | 0.002(0.925) | 0.238(0.000) | – | – |
| | Model 3 | 0.125(0.000) | 0.038(0.429) | 0.019(0.135) | 0.128(0.000) | 0.012(0.382) | 0.118(0.000) | 0.666(0.000) | 1.952(0.000) |
| RAVLT.forgetting | Model 1 | −0.026(0.580) | 0.002(0.979) | −0.043(0.056) | −0.297(0.000) | −0.158(0.000) | – | – | – |
| | Model 2 | −0.072(0.055) | −0.231(0.001) | −0.033(0.075) | −0.256(0.000) | −0.161(0.000) | 0.148(0.000) | – | – |
| | Model 3 | 0.062(0.219) | −0.120(0.202) | 0.002(0.953) | −0.173(0.008) | −0.047(0.064) | 0.024(0.353) | 0.637(0.000) | 0.539(0.000) |
| RAVLT.perc.forgetting | Model 1 | −0.106(0.005) | 0.120(0.081) | −0.149(0.000) | −0.019(0.685) | −0.045(0.018) | – | – | – |
| | Model 2 | −0.107(0.001) | 0.295(0.000) | −0.125(0.000) | −0.009(0.830) | −0.047(0.003) | 0.188(0.000) | – | – |
| | Model 3 | −0.054(0.052) | 0.022(0.659) | −0.054(0.000) | −0.170(0.000) | 0.003(0.855) | 0.093(0.000) | 0.989(0.000) | 1.481(0.000) |

TABLE 4
(*Continued*)

| Score | Model | Gender | Handedness | Education | Retirement | Age | APOE4 | MCI | AD |
|---|---|---|---|---|---|---|---|---|---|
| TRABSCOR | Model 1 | −0.036(0.436) | −0.088(0.293) | −0.230(0.000) | −0.021(0.712) | 0.003(0.876) | – | – | – |
| | Model 2 | 0.004(0.926) | −0.131(0.076) | −0.182(0.000) | 0.023(0.065) | 0.042(0.037) | 0.180(0.000) | – | – |
| | Model 3 | 0.010(0.803) | −0.012(0.867) | −0.143(0.000) | 0.000(1.000) | 0.061(0.002) | 0.052(0.011) | 0.642(0.000) | 1.484(0.000) |
| DIGITSCOR | Model 1 | 0.189(0.000) | 0.038(0.584) | 0.227(0.000) | 0.151(0.002) | −0.076(0.000) | – | – | – |
| | Model 2 | 0.253(0.000) | −0.118(0.057) | 0.192(0.000) | −0.051(0.230) | −0.088(0.000) | −0.176(0.000) | – | – |
| | Model 3 | 0.238(0.000) | −0.170(0.010) | 0.130(0.000) | 0.021(0.639) | −0.052(0.004) | −0.036(0.058) | −0.512(0.000) | −1.389(0.000) |
| LDELTOTAL | Model 1 | 0.077(0.067) | −0.209(0.009) | 0.187(0.000) | −0.024(0.659) | −0.004(0.846) | – | – | – |
| | Model 2 | 0.293(0.000) | −0.159(0.006) | 0.199(0.000) | 0.100(0.010) | −0.035(0.025) | −0.349(0.000) | – | – |
| | Model 3 | 0.021(0.396) | −0.059(0.189) | 0.105(0.000) | −0.052(0.090) | −0.034(0.007) | −0.135(0.000) | −1.329(0.000) | −1.868(0.000) |
| MMSE | Model 1 | −0.043(0.215) | 0.302(0.00) | 0.071(0.000) | −0.121(0.008) | −0.049(0.006) | – | – | – |
| | Model 2 | −0.006(0.847) | 0.240(0.000) | 0.141(0.000) | −0.161(0.000) | −0.008(0.609) | −0.230(0.000) | – | – |
| | Model 3 | −0.159(0.000) | 0.146(0.002) | 0.019(0.130) | −0.080(0.011) | −0.038(0.003) | −0.103(0.000) | −0.670(0.000) | −1.752(0.000) |
| RAVLT.learning | Model 1 | 0.283(0.000) | −0.102(0.222) | 0.123(0.000) | 0.048(0.401) | −0.031(0.158) | – | – | – |
| | Model 2 | 0.268(0.000) | 0.092(0.018) | 0.124(0.000) | −0.002(0.972) | −0.047(0.010) | −0.227(0.000) | – | – |
| | Model 3 | 0.139(0.000) | 0.122(0.047) | 0.079(0.000) | −0.059(0.160) | 0.018(0.296) | −0.045(0.009) | −0.910(0.000) | −1.240(0.000) |
| RAVLT.immediate | Model 1 | 0.425(0.000) | 0.169(0.050) | 0.228(0.000) | 0.022(0.708) | −0.049(0.033) | – | – | – |
| | Model 2 | 0.457(0.000) | 0.094(0.173) | 0.227(0.000) | 0.001(0.980) | −0.039(0.034) | −0.268(0.000) | – | – |
| | Model 3 | 0.287(0.000) | −0.008(0.855) | 0.147(0.000) | −0.134(0.000) | −0.008(0.531) | −0.106(0.000) | −0.973(0.000) | −1.602(0.000) |

Model 1 corrects for all covariates, except APOE4 and the baseline disease status; Model 2 corrects for all covariates, except the baseline disease status, and Model 3 corrects for all covariates



(a)

(b)

FIG. 6. *Panel* (*a*) *presents positions of the important SNPs for the* 13 *scores for Model* 1. *The colors represent the* 13 *scores*, *and each signal dot indicates the SNP is selected for the corresponding score. Panel* (*b*) *gives* $p$*-values of the selected SNPs for Model* 1. *The dashed line indicates the threshold* $p$*-value*$= 0.05/10{,}154$.

on chromosome 19 are found to be important for all 13 cognitive scores. Table 5 presents the SNPs on chromosome 19 that are identified as important for at least three cognitive scores. The rs429358 SNP on chromosome 19, which is one of the two variants for the well-known APOE alleles, is significant for ADAS11, ADAS13, FAQ, MMSE, RAVLT.immediate, and RAVLT.perc.forgetting. Other notable SNPs include rs283812 in the PVRL2 region, rs769449 in the APOE region, and rs66626994 in the APOC1 region. The APOE, PVRL2, and APOC1 regions within the cytogenetic region 19q13.32 are considered high AD-risk regions, as supported by previous research (Carrasquillo et al. (2009), Vermunt et al. (2019)).

In addition to chromosome 19, two SNPs are identified as important for at least three cognitive scores from other chromosomes. These include rs28414114 from chromosome 14, with the smallest $p$-value of 7e−14, and rs12108758 from chromosome 5, with the smallest $p$-

TABLE 5
*Detailed information of the common SNPs from the* 19*th chromosome for at least three scores for Model* 1

| SNP | Chr | Base Pair | Scores |
|---|---|---|---|
| rs283812 | 19 | 45388568 | CDRSB, LDELTOTAL, MMSE |
| rs769449 | 19 | 45410002 | LDELTOTAL, MMSE, RAVLT.immediate |
| rs429358 | 19 | 45411941 | ADAS11, ADAS13, FAQ, MMSE, RAVLT.immediate, RAVLT.perc.forgetting |
| rs66626994 | 19 | 45428234 | ADAS13, LDELTOTAL, RAVLT.immediate |

value of 2e−9. Moreover, the $p$-values of the selected SNPs can be found in Figure 6(b). The $p$-values for several key SNPs are smaller than $0.05/10{,}145$, which represents the threshold for significance after adjusting for multiple testing (with 10,145 SNPs remaining after screening).

5.2. *Conditional GIC* (*CGIC*) *pathways given APOE*4. To address question (Q2), we fit Model 2 by explicitly accounting for APOE4 and exclude the SNPs in the 19q13.32 region from the candidate SNPs. We apply the same screening step and Algorithm 1 to Model 2 for each cognitive score. In our dataset 230 subjects had one APOE4 allele, and 67 subjects had two APOE4 alleles. The cytogenetic region 19q13.32 comprises 6376 SNPs, including the well-known APOE (Bertram and Tanzi (2012)). This information enables us to better understand the conditional effects of other SNPs on cognitive scores, given the presence of APOE4 alleles.

Table 4 presents the estimation results corresponding to Model 2. Estimates of the demographic covariates in Model 2 are similar to their corresponding estimates in Model 1. The number of APOE4 alleles is significant for all cognitive scores, showing negative effects on cognitive ability. Estimates of the hippocampal surface for the 13 cognitive scores in Model 2 are also similar to those in Model 1; therefore, we provide them in the Supplementary Material (Li et al. (2024)).

Figure S5(a) in the Supplementary Material (Li et al. (2024)) displays the ideogram of the selected important SNPs for Model 2. For each cognitive score, the selected significant SNPs in Model 2 share some similarities with those in Model 1. For example, rs28414114 from chromosome 14 is identified as important for both ADAS11 and ADAS13 in both models. Additionally, the regions of the selected SNPs for some cognitive scores are similar between the two models. For instance, the positions of the important SNPs for TABSCOR in chromosome 6 in Model 1 range from 81858848 to 91078328 with the smallest $p$-value of 9.06e−11, and from 120546293 to 120765041 with the smallest $p$-value of 8e−9. In Model 2 the important positions range from 80949537 to 94053838 with the smallest $p$-value of 1.16e−9, and from 120533534 to 120771669 with the smallest $p$-value of 6e−15.

5.3. *CGIC pathways given APOE*4 *and disease status*. To address (Q3), we fit Model 3 by explicitly accounting for both APOE4 and baseline diagnosis information. The baseline diagnosis status is represented by two dummy variables: MCI and AD. Since clinical notes offer supplementary information and are considered on a case-by-case basis, the effects of SNPs on changes in cognitive performance might be confounded with the effects of differences in baseline diagnosis. We are interested in determining whether the relationships would change when adjusting for baseline diagnosis status. We apply the same screening step and Algorithm 1 to Model 3 for each cognitive score.

Table 4 also presents the related estimation results corresponding to Model 3. After accounting for the baseline diagnosis status, almost all estimates of demographic covariates and APOE4 in Model 3 are smaller than their corresponding estimates in Models

1 and 2. The baseline MCI status has significant positive effects on ADAS11, ADAS13, CDRSB, FAQ, RAVLT.forgetting, RAVLT.perc.forgetting, and TRABSCOR, while exhibiting significant negative effects on DIGITSCOR, LDELTOTAL, MMSE, RAVLT.learning, and RAVLT.immediate. The baseline AD status generally has stronger effects on the 13 cognitive scores at Month 12 than the baseline MCI status. Similar patterns of hippocampal estimates are also observed for the 13 cognitive scores, as seen in Section 5.1, and the corresponding results are included in the Supplementary Material (Li et al. (2024)).

Figure S5(b) in the Supplementary Material (Li et al. (2024)) displays the ideogram of the selected important SNPs for Model 3. The selected important SNPs appear quite different from the SNPs in Figure 6(a). This is reasonable, as we consider the baseline diagnosis status in the screening step and consistently include it in the model. The selected important SNPs for at least three cognitive scores are rs13101604 from chromosome 4 with the smallest $p$-value of 4.18e$-$9, rs2442696 from chromosome 4 with the smallest $p$-value of 6.96e$-$11, and rs4761161 from chromosome 12 with the smallest $p$-value of 2.11e$-$09.

5.4. *Comparisons of the three models.* In this subsection we further address (Q2) and (Q3) by comparing the three models in terms of shared and different heritability patterns for the 13 cognitive scores and the proportions of variations explained in cognitive deficits by the three types of data: genetic data, controlling covariates, and hippocampal surface data. The average computation times of the proposed method are 2.25 hours with a standard error of 0.41 hours for Model 1, 2.31 hours with a standard error of 0.44 hours for Model 2, and 2.19 hours with a standard error of 0.45 hours for Model 3.

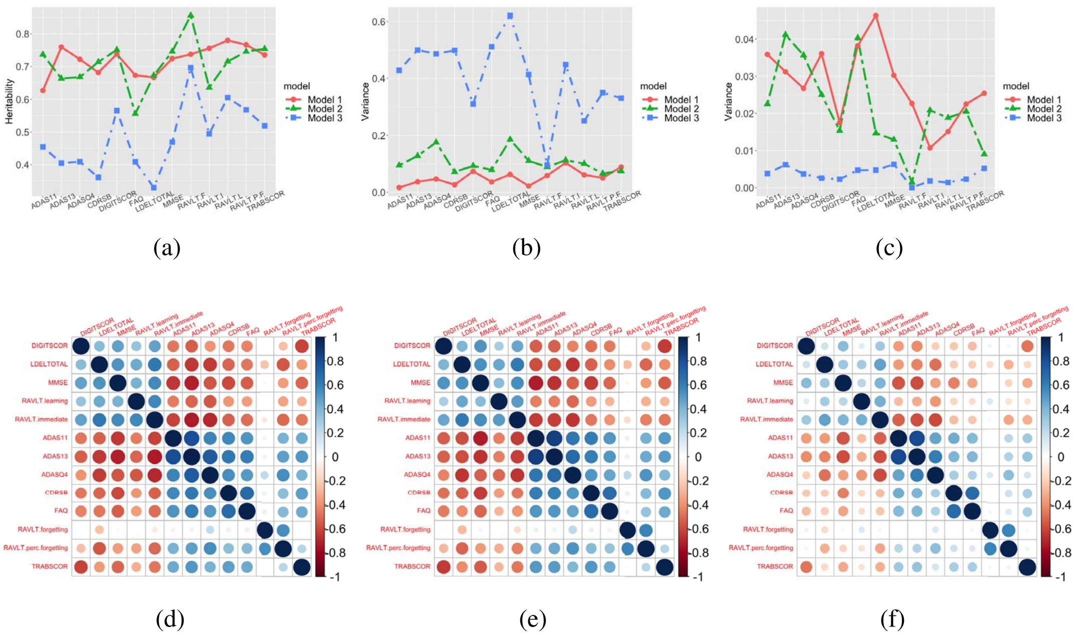
Although most human traits have a polygenic architecture (Wray et al. (2018)), heritability can be used to measure how much of the variation in each score is due to variation in genetic data (Yang et al. (2010)). By definition, we estimate the heritability for the three models by calculating the phenotypic variance due to the genetic variables. We estimate the phenotypic variance by the empirical variance of $X_{iG}^{\top}\hat{\beta}_G$, where $X_{iG}$ is the genetic variables of the $i$th subject and $\hat{\beta}_G$ is the corresponding coefficient estimates. For Model 2 and Model 3, estimates of the heritability are based on the considered SNPs, which excludes the SNPs in the cytogenetic region 19q13.32.

Figure 7(a) presents the heritability estimates of the genetic effects for the cognitive scores. Heritability of the cognitive scores is estimated to be $62.69\% \sim 78.01\%$ for Model 1, $55.56\% \sim 85.62\%$ for Model 2, and $33.02\% \sim 69.66\%$ for Model 3. The remaining heritability of the scores, especially RAVLT.forgetting, is still relatively high even after accounting for APOE4. This is consistent with previous research and suggests that memory functioning in AD is under strong genetic influence that is only partly attributable to the APOE genotype (Wilson et al. (2011)). However, there are $1.1\% \sim 35.3\%$ decreases in the heritability estimates of the 13 cognitive scores for Model 3. This reveals that the baseline diagnosis status explains a part of the cognitive function associated with the polygenic effect.

We also examine the effect size of the controlling covariates by calculating the proportion of variance explained by these covariates in Figure 7(b). The proportions of variance explained by the controlling variables increase with the inclusion of the number of APOE4 alleles and the baseline disease status.

Figure 7(c) illustrates the effect size of imaging covariates. The hippocampal surface data account for approximately 1% to 4.6% of the total variations in 13 cognitive scores for Model 1, 0.1% to 4.1% for Model 2, and 0.005% to 0.63% for Model 3. These findings indicate that the baseline diagnosis status explains a more substantial portion of cognitive function associated with hippocampal data than the number of APOE4 gene alleles.

We also compare the effects of genetic data with those of imaging data. The results suggest that cognitive function may have a polygenic inheritance, meaning it is not controlled by a



(a) (b) (c)

(d) (e) (f)

FIG. 7. *Heritability estimates of the genetic data* (*a*), *the proportions of variance explained by the controlling variables* (*b*), *and the imaging data* (*c*). *Genetic correlations between the* 13 *cognitive scores for Model* 1 (*d*), *Model* 2 (*e*), *and Model* 3 (*f*). *The line with points represents Model* 1, *the line with triangles represents Model* 2, *and the line with squares represents Model* 3.

single gene but multiple genes that each contribute a small portion to the overall outcome. On average, 75 genes are selected as important for the cognitive scores in the three models. The variance, explained by a single gene, ranges from 0.84% to 1.05% in Model 1, 0.74% to 1.14% in Model 2, and 0.44% to 0.93% in Model 3, which is comparable to that of imaging data.

In addition, we calculate the estimated $p$-values of left and right hippocampal surface data for the 13 cognitive scores in the three models, as shown in Table 6. The left hippocampus is significant at a 5% significance level for nine scores in Model 1, 11 scores in Model 2, and five scores in Model 3. The right hippocampus is significant for seven scores in Model 1, 11 scores in Model 2, and three scores in Model 3. Both the left and right hippocampi are significant for ADAS13 and MMSE, which is consistent with findings from Morrison et al. (2022) and Peng et al. (2015). It is also noteworthy that most $p$-values in Model 3 are larger than the corresponding $p$-values in Models 1 and 2.

Genetic correlation has been proposed to describe the shared genetic associations within pairs of quantitative traits, and it can be calculated as the correlation of the genetic effects of numerous SNPs on the traits (Zhao and Zhu (2022)). Providing crucial information about fundamental biological pathways and describing the shared genetic etiology of the 13 scores, we calculate the genetic correlations between the 13 scores for the three models we considered in Figures 7(d)–7(f).

There are strong genetic correlations between the 13 scores for Model 1, suggesting an overall similarity of the genetic architecture on brain functions in Month 12 characterized by these scores. The genetic correlations adjusted for the number of APOE4 alleles are similar to those for Model 1, albeit with slightly smaller values, indicating shared genetic effects for the 13 scores besides the well-known APOE4 gene effect. However, the genetic correlations significantly decrease when additionally controlling for the baseline diagnosis status, which is assumed to explain a large part of the shared genetic effect on the cognitive scores.

TABLE 6
*Estimated p-values of hippocampal surface data for the* 13 *cognitive scores*

| | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| Score | Left | Right | Left | Right | Left | Right |
| ADAS11 | 3.66E−05 | 1.35E−08 | 6.72E−05 | 1.04E−05 | 0.054 | 0.046 |
| ADAS13 | 1.58E−08 | 1.49E−09 | 6.36E−08 | 2.11E−05 | 0.016 | 0.023 |
| ADASQ4 | 2.55E−07 | 2.82E−05 | 3.41E−07 | 2.88E−06 | 0.052 | 0.096 |
| CDRSB | 4.56E−07 | 4.27E−06 | 2.25E−05 | 0.008 | 0.105 | 0.142 |
| FAQ | 3.72E−05 | 0.753 | 0.998 | 0.003 | 0.039 | 0.102 |
| RAVLT.forgetting | 0.707 | 0.999 | 0.160 | 0.257 | 0.999 | 0.120 |
| RAVLT.perc.forgetting | 0.001 | 0.824 | 9.89E−05 | 2E−04 | 0.132 | 0.071 |
| TRABSCOR | 0.916 | 0.999 | 0.007 | 0.001 | 0.011 | 0.090 |
| DIGITSCOR | 0.036 | 0.999 | 0.001 | 0.381 | 0.117 | 0.582 |
| LDELTOTAL | 2.19E−07 | 1.82E−09 | 0.001 | 7.22E−06 | 0.032 | 0.160 |
| MMSE | 2.66E−10 | 5.35E−11 | 0.002 | 1.06E−05 | 0.022 | 0.039 |
| RAVLT.learning | 0.998 | 7.80E−05 | 2E−04 | 5.70E−05 | 0.999 | 0.080 |
| RAVLT.immediate | 0.999 | 0.999 | 7.83E−05 | 4E−04 | 0.161 | 0.329 |

Model 1 corrects for all covariates, except APOE4 and the baseline disease status; Model 2 corrects for all covariates, except the baseline disease status, and Model 3 corrects for all covariates.

Moreover, this reveals greater genetic heterogeneity after accounting for the baseline diagnosis status. One potential reason is that the AD population is genetically heterogeneous (Lo et al. (2019)). Adjusting for disease status suggests that, within a given diagnostic group, substantial variation in cognitive scores is explained by genetic markers, indicating heterogeneity within disease classifications partially explained by genetics. This may imply that different therapies are needed for different symptoms after AD onset.

**6. Discussion.** This paper aims to map the biological pathways of phenotypes of interest from the ADNI study by integrating GIC data. The high-dimensional genetic data and the features of the hippocampal surface data motivate us to consider a high-dimensional PFLR to establish the associations between the genetic and imaging data with the phenotype of interest. We propose a new estimation method for high-dimensional PFLR, under the RKHS framework and the $\ell_0$ penalty, and investigate the theoretical results of the estimators. Through the analyses of the ADNI study, we demonstrate that the proposed method is a valuable statistical tool for quantifying the complex relationships between phenotypes of interest and the GIC data.

Although the proposed method considers one functional covariate, it can be extended to accommodate multiple functional covariates. We can consider the following high-dimensional PFLR with multiple functional predictors:

$$Y_i = \alpha + X_i^\top \beta + \sum_{k=1}^{K} \int_{\mathcal{T}} Z_{ik}(t)\xi_k(t)\,dt + \epsilon_i.$$

Each $\xi_k(t)$ is assumed to be in a reproducing kernel Hilbert space. Similar to the minimization problem in (2) that imposes $\ell_0$ penalty to $\beta$ and RKHS roughness penalty on $\xi(t)$, we can estimate $\{\xi_k(t)\}$s by imposing RKHS roughness penalty on each $\xi_k(t)$. The representer theorem and the theoretical results can be obtained similarly. Such extensions are worthy of further investigation.

**Acknowledgments.** These authors contributed equally: Ting Li and Yang Yu.

The authors would like to thank the referees, the Associate Editor, and the Editor for their constructive comments that improved the quality of this paper.

Hongtu Zhu is also affiliated with Department of Statistics and Operations Research, Departments of Genetics, and Departments of Computer Science, University of North Carolina at Chapel Hill.

Data used in 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.

**Funding.** Research reported in this publication was partially supported by the National Institute On Aging of the National Institutes of Health under Award Number RF1AG082938. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## SUPPLEMENTARY MATERIAL

**Additional results and proofs** (DOI: 10.1214/23-AOAS1808SUPPA; .pdf). The Supplementary Material contains additional real data analysis, additional simulation results, theoretical properties of the estimators, lemmas for the purely functional linear model, auxiliary lemmas, and details of all the proofs.

**Data and code** (DOI: 10.1214/23-AOAS1808SUPPB; .zip). The data and implementation R code, along with its documentation, are available as an online supplement (https://github.com/BIG-S2/PFLR_RKHS_code).

## REFERENCES

BATTISTA, P., SALVATORE, C. and CASTIGLIONI, I. (2017). Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study. *Behav. Neurol.* **2017** 1–19.

BERTRAM, L. and TANZI, R. E. (2012). The genetics of Alzheimer's disease. *Prog. Mol. Biol. Transl. Sci.* **107** 79–100.

BRAAK, H. and BRAAK, E. (1998). *Evolution of Neuronal Changes in the Course of Alzheimer's Disease*. Springer, Berlin.

CAI, T. T. and YUAN, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Statist.* **39** 2330–2355.

CAI, T. T. and YUAN, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.* **107** 1201–1216.

CARRASQUILLO, M. M., ZOU, F., PANKRATZ, V. S., WILCOX, S. L., MA, L., WALKER, L. P., YOUNKIN, S. G., YOUNKIN, C. S., YOUNKIN, L. H. et al. (2009). Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat. Genet.* **41** 192–198.

CRUCIANI, F., ALTMANN, A., LORENZI, M., MENEGAZ, G. and GALAZZO, I. B. (2022). What PLS can still do for imaging genetics in Alzheimer's disease. In 2022 *IEEE-EMBS International Conference on Biomedical and Health Informatics* (*BHI*) 1–4. IEEE.

DE AQUINO, C. H. (2021). Methodological issues in randomized clinical trials for prodromal Alzheimer's and Parkinson's disease. *Front. Neurol.* **12** 694329.

DE FLORES, R., JOIE, R. L. and CHÉTELAT, G. (2015). Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease. *Neuroscience* **309** 29–50.

DEARY, I. J., COX, S. R. and HILL, W. D. (2022). Genetic variation, brain, and intelligence differences. *Mol. Psychiatry* **27** 335–353.

DUKART, J., SAMBATARO, F. and BERTOLINO, A. (2016). Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers. *J. Alzheimer's Dis.* **49** 1143–1159.

DUONG, M. T., DAS, S. R., LYU, X., XIE, L., RICHARDSON, H., XIE, S. X., YUSHKEVICH, P. A., WOLK, D. A. and NASRALLAH, I. M. (2022). Dissociation of tau pathology and neuronal hypometabolism within the ATN framework of Alzheimer's disease. *Nat. Commun.* **13** 1–15.

ELLIOTT, L. T., SHARP, K., ALFARO-ALMAGRO, F., SHI, S., MILLER, K. L., DOUAUD, G., MARCHINI, J. and SMITH, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK biobank. *Nature* **562** 210–216.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911.

FRISONI, G. B., GANZOLA, R., CANU, E., RÜB, U., PIZZINI, F. B., ALESSANDRINI, F., ZOCCATELLI, G., BELTRAMELLO, A., CALTAGIRONE, C. et al. (2008). Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 Tesla. *Brain* **131** 3266–3276.

GE, T., NICHOLS, T. E., LEE, P. H., HOLMES, A. J., ROFFMAN, J. L., BUCKNER, R. L., SABUNCU, M. R. and SMOLLER, J. W. (2015). Massively expedited genome-wide heritability analysis (MEGHA). *Proc. Natl. Acad. Sci. USA* **112** 2479–2484.

GRASSI, M., ROULEAUX, N., CALDIROLA, D., LOEWENSTEIN, D., SCHRUERS, K., PERNA, G., DUMONTIER, M. and INITIATIVE, A. D. N. (2019). A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures. *Front. Neurol.* 756.

GUERREIRO, R. and BRAS, J. (2015). The age factor in Alzheimer's disease. *Gen. Med.* **7** 1–3.

HOBART, J., CANO, S., POSNER, H., SELNES, O., STERN, Y., THOMAS, R., ZAJICEK, J. and INITIATIVE, A. D. N. (2013). Putting the Alzheimer's cognitive test to the test I: Traditional psychometric methods. *Alzheimer's Dement.* **9** S4–S9.

HUANG, C., THOMPSON, P., WANG, Y., YU, Y., ZHANG, J., KONG, D., COLEN, R. R., KNICKMEYER, R. C., ZHU, H. et al. (2017). FGWAS: Functional genome wide association analysis. *NeuroImage* **159** 107–121.

HUANG, J., JIAO, Y., LIU, Y. and LU, X. (2018). A constructive approach to $L_0$ penalized regression. *J. Mach. Learn. Res.* **19** 403–439.

JACK JR., C. R., KNOPMAN, D. S., JAGUST, W. J., PETERSEN, R. C., WEINER, M. W., AISEN, P. S., SHAW, L. M., VEMURI, P., WISTE, H. J. et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* **12** 207–216.

JACK JR., C. R., KNOPMAN, D. S., JAGUST, W. J., SHAW, L. M., AISEN, P. S., WEINER, M. W., PETERSEN, R. C. and TROJANOWSKI, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9** 119–128.

KNUTSON, K. A., DENG, Y. and PAN, W. (2020). Implicating causal brain imaging endophenotypes in Alzheimer's disease using multivariable IWAS and GWAS summary data. *NeuroImage* **223** 117347.

KONG, D., XUE, K., YAO, F. and ZHANG, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103** 147–159.

LI, S., SHI, F., PU, F., LI, X., JIANG, T., XIE, S. and WANG, Y. (2007). Hippocampal shape analysis of Alzheimer disease based on machine learning methods. *Am. J. Neuroradiol.* **28** 1339–1345.

LI, T., YU, Y., MARRON, J. and ZHU, H. (2024). Supplement to "A partially functional linear regression framework for integrating genetic, imaging, and clinical data." https://doi.org/10.1214/23-AOAS1808SUPPA, https://doi.org/10.1214/23-AOAS1808SUPPB

LI, T. and ZHU, Z. (2020). Inference for generalized partial functional linear regression. *Statist. Sinica* **30** 1379–1397.

LIN, D., CAO, H., CALHOUN, V. D. and WANG, Y. P. (2014). Sparse models for correlative and integrative analysis of imaging and genetic data. *J. Neurosci. Methods* **237** 69–78.

LO, M.-T., KAUPPI, K., FAN, C.-C., SANYAL, N., REAS, E. T., SUNDAR, V., LEE, W.-C., DESIKAN, R. S., MCEVOY, L. K. et al. (2019). Identification of genetic heterogeneity of Alzheimer's disease across age. *Neurobiol. Aging* **84** 243.e1–243.e9.

MA, H., LI, T., ZHU, H. and ZHU, Z. (2019). Quantile regression for functional partially linear model in ultrahigh dimensions. *Comput. Statist. Data Anal.* **129** 135–147.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr., C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J. et al. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **7** 263–269.

Morrison, C., Dadar, M., Shafiee, N., Villeneuve, S., Collins, D. L., Initiative, A. D. N. et al. (2022). Regional brain atrophy and cognitive decline depend on definition of subjective cognitive decline. *NeuroImage Clin.* **33** 102923.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack Jr., C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W. and Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's Dement.* **1** 55–66.

Nathoo, F. S., Kong, L. and Zhu, H. (2019). A review of statistical methods in imaging genetics. *Canad. J. Statist.* **47** 108–131.

Ossenkoppele, R., Leuzy, A., Cho, H., Sudre, C. H., Strandberg, O., Smith, R., Palmqvist, S., Mattsson-Carlgren, N., Olsson, T. et al. (2021). The impact of demographic, clinical, genetic, and imaging variables on tau PET status. *Eur. J. Nucl. Med. Mol. Imaging* **48** 2245–2258.

Pedraza, O., Bowers, D. and Gilmore, R. (2004). Asymmetry of the hippocampus and amygdala in MRI volumetric measurements of normal adults. *J. Int. Neuropsychol. Soc.* **10** 664–678.

Peng, G.-P., Feng, Z., He, F.-P., Chen, Z.-Q., Liu, X.-Y., Liu, P. and Luo, B.-Y. (2015). Correlation of hippocampal volume and cognitive performances in patients with either mild cognitive impairment or Alzheimer's disease. *CNS Neuroscience Ther.* **21** 15–22.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York.

Selkoe, D. J. and Hardy, J. (2016). The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8** 595–608.

Shen, L. and Thompson, P. M. (2020). Brain imaging genomics: Integrated analysis and machine learning. *Proc. IEEE Inst. Electr. Electron. Eng.* **108** 125–162.

Smith, S. M., Douaud, G., Chen, W., Hanayik, T., Alfaro-Almagro, F., Sharp, K. and Elliott, L. T. (2021). An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat. Neurosci.* **24** 737–745.

Thompson, P. M., Jahanshad, N., Ching, C. R., Salminen, L. E., Thomopoulos, S. I., Bright, J., Baune, B. T., Bertolín, S., Bralten, J. et al. (2020). ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl. Psychiatry* **10** 1–28.

Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., DeCarli, C., Green, R. C., Harvey, D., Jack Jr., C. R., Jagust, W. et al. (2021). Using the Alzheimer's disease neuroimaging initiative to improve early detection, diagnosis, and treatment of Alzheimer's disease. *Alzheimer's Dement.* **18** 824–857.

Vermunt, L., Sikkes, S. A., Van Den Hout, A., Handels, R., Bos, I., Van Der Flier, W. M., Kern, S., Ousset, P.-J., Maruff, P. et al. (2019). Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in relation to age, sex, and APOE genotype. *Alzheimer's Dement.* **15** 888–898.

Vina, J. and Lloret, A. (2010). Why women have more Alzheimer's disease than men: Gender and mitochondrial toxicity of amyloid-$\beta$ peptide. *J. Alzheimer's Dis.* **20** S527–S533.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

Wang, L., Kim, Y. and Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536.

Wilson, R. S., Barral, S., Lee, J. H., Leurgans, S. E., Foroud, T. M., Sweet, R. A., Graff-Radford, N., Bird, T. D., Mayeux, R. et al. (2011). Heritability of different forms of memory in the Late Onset Alzheimer's Disease Family Study. *J. Alzheimer's Dis.* **23** 249–255.

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. and Visscher, P. M. (2018). Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173** 1573–1580.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.

Yao, F., Sue-Chee, S. and Wang, F. (2017). Regularized partially functional quantile regression. *J. Multivariate Anal.* **156** 39–56.

Yu, D., Wang, L., Kong, D. and Zhu, H. (2022). Mapping the genetic-imaging-clinical pathway with applications to Alzheimer's disease. *J. Amer. Statist. Assoc.* **117** 1656–1668.

YUAN, M. and CAI, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444.

ZHAO, B., LI, T., SMITH, S. M., XIONG, D., WANG, X., YANG, Y., LUO, T., ZHU, Z., SHAN, Y. et al. (2022). Common variants contribute to intrinsic human brain functional networks. *Nat. Genet.* **54** 508–517.

ZHAO, B., LI, T., YANG, Y., WANG, X., LUO, T., SHAN, Y., ZHU, Z., XIONG, D., HAUBERG, M. E. et al. (2021). Common genetic variation influencing human white matter microstructure. *Science* **372** eabf3736.

ZHAO, B., LUO, T., LI, T., LI, Y., ZHANG, J., SHAN, Y., WANG, X., YANG, L., ZHOU, F. et al. (2019). Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat. Genet.* **51** 1637–1644.

ZHAO, B. and ZHU, H. (2022). On genetic correlation estimation with summary statistics from genome-wide association studies. *J. Amer. Statist. Assoc.* **117** 1–11.

ZHU, H., LI, T. and ZHAO, B. (2023). Statistical learning methods for neuroimaging data analysis with applications. *Annu. Rev. Biomed. Data Sci.* **6** 73–104.