Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Sparse Learning and Structure Identification for Ultrahigh-Dimensional Image-on-Scalar Regression

Xinyi Li, Li Wang, Huixia Judy Wang & for the Alzheimer's Disease Neuroimaging Initiative

To cite this article: Xinyi Li, Li Wang, Huixia Judy Wang & for the Alzheimer's Disease Neuroimaging Initiative (2020): Sparse Learning and Structure Identification for Ultrahigh-Dimensional Image-on-Scalar Regression, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1753523

To link to this article: https://doi.org/10.1080/01621459.2020.1753523

Published online: 26 May 2020.

Submit your article to this journal

Article views: 777

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–15: Theory and Methods
https://doi.org/10.1080/01621459.2020.1753523

Taylor & Francis
Taylor & Francis Group

Check for updates

Sparse Learning and Structure Identification for Ultrahigh-Dimensional Image-on-Scalar Regression

Xinyi Li[a,b], Li Wang[c], Huixia Judy Wang[d], and for the Alzheimer's Disease Neuroimaging Initiative

[a]Statistical and Applied Mathematical Sciences Institute (SAMSI), Durham, NC; [b]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC; [c]Department of Statistics, Iowa State University, Ames, IA; [d]Department of Statistics, George Washington University, Washington, DC

ABSTRACT

This article considers high-dimensional image-on-scalar regression, where the spatial heterogeneity of covariate effects on imaging responses is investigated via a flexible partially linear spatially varying coefficient model. To tackle the challenges of spatial smoothing over the imaging response's complex domain consisting of regions of interest, we approximate the spatially varying coefficient functions via bivariate spline functions over triangulation. We first study estimation when the active constant coefficients and varying coefficient functions are known in advance. We then further develop a unified approach for simultaneous sparse learning and model structure identification in the presence of ultrahigh-dimensional covariates. Our method can identify zero, nonzero constant, and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent and asymptotically normal for constant coefficient estimators. The method is evaluated by Monte Carlo simulation studies and applied to a dataset provided by the Alzheimer's Disease Neuroimaging Initiative. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2019
Accepted April 2020

KEYWORDS
Bivariate splines; Imaging data; Triangulation; Varying coefficient models

1. Introduction

In recent years, there has been explosive growth in the number of imaging studies in medical research, thus, boosting the investigation of "next-generation functional data" that includes images and shapes (Wang, Chiou, and Muller 2016). Functional data analysis (FDA) methods have been widely studied in the literature, but most focus has been placed on one-dimensional curves and functional data distributed on rectangular domains; see Morris (2015), Ramsay and Silverman (2005), Wang, Chiou, and Muller (2016), and references therein for a comprehensive review of FDA. Little has been done on the development of advanced statistical methods to analyze multidimensional functional data (such as two-dimensional or three-dimensional) or functional data over complex domains.

In this article, we consider image-on-scalar regression with imaging responses in a template and scalar predictors from n subjects. Let Ω represent a domain of arbitrary shape of a complex object, and let $\boldsymbol{s} = (s_1, s_2)^\top$ denote a point in Ω. Without loss of generality, in what follows, Ω is assumed to be a compact set in $\mathbb{R}^2$. For subject $i = 1, \ldots, n$, we observe a real-valued imaging measure $Y_i(\boldsymbol{s}_j)$ at point $\boldsymbol{s}_j \in \Omega$ ($j = 1, \ldots, N_s$) and scalar predictors X_{ik}, $k = 1, \ldots, p_n$, such as age, gender, height, environmental and genetic factors. For notation simplicity, we write p instead of p_n in the rest of the article, even though the number of scalar predictors may grow with the number of subjects. To characterize the relationship between imaging responses and scalar covariates, we consider the spatially varying coefficient model (SVCM)

$$Y_i(\boldsymbol{s}) = \mu_0(\boldsymbol{s}) + \sum_{k=1}^{p} X_{ik} f_k(\boldsymbol{s}) + \eta_i(\boldsymbol{s}) + \varepsilon_i(\boldsymbol{s}),$$
$$i = 1, \ldots, n, \ \boldsymbol{s} \in \Omega, \tag{1}$$

where μ_0 and $f_1, \ldots, f_p$ are unknown smooth functions; $\eta_1, \ldots, \eta_n$ characterize individual image variations, and are independent and identical copies of a stochastic process with mean zero and covariance function $G_\eta(\boldsymbol{s}, \boldsymbol{s}')$; $\varepsilon_1(\boldsymbol{s}), \ldots, \varepsilon_n(\boldsymbol{s})$ are measurement errors, and independent and identical copies of a random process with mean zero, and covariance $\mathrm{cov}\{\varepsilon_i(\boldsymbol{s}), \varepsilon_i(\boldsymbol{s}')\} = \sigma^2(\boldsymbol{s}) I(\boldsymbol{s} = \boldsymbol{s}')$. Each η_i is independent of ε_i, and each η_i and ε_i are independent of X_{ik}'s. In the following, we assume both X_{ik}'s and $Y_i(\boldsymbol{s})$'s are centered by subtracting their mean, so that $\mathrm{E}X_{ik} = 0$ for all $k = 1, \ldots, p$ and $\mathrm{E}Y_i(\boldsymbol{s}) = 0$ for all $\boldsymbol{s} \in \Omega$.

The SVCM in (1) is powerful for modeling the nonstationarity of regression coefficients over space. Under this model, the imaging response is associated with scalar covariates through functional linear regression, but the regression coefficients can vary from location to location across the response image and are modeled as a nonparametric function of spatial coordinates. Model (1) assumes that all the coefficients are spatially varying. While in reality, some covariates may have homogenous effects while others have heterogeneous effects across locations. If we

CONTACT Li Wang lilywang@iastate.edu Department of Statistics, Iowa State University, Ames, IA 50011.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2020 American Statistical Association

can correctly identify the constant coefficients, it is possible to improve the estimation efficiency of the SVCM by considering a partially linear spatially varying coefficient model (PLSVCM), a more parsimonious special case of the SVCM; see model (2) in Section 2.

Our research on the SVCM and PLSVCM is related to works on varying coefficient models (VCMs; see Hastie and Tibshirani 1993; Huang, Wu, and Zhou 2004; Tang et al. 2012; Jiang et al. 2013; Gu et al. 2014; Liu, Li, and Wu 2014) and partially linear VCMs (PLVCMs; see Ahmad, Leelahanon, and Li 2005; Fan and Huang 2005; Wang, Zhu, and Zhou 2009). While there is a growing body of literature on statistical tools for univariate smoothing in which the index variable is a scalar such as time, there are far fewer works on how to estimate SVCMs where the index variable $\boldsymbol{s}$ is two-dimensional. In a spatial regression setting, Gelfand et al. (2003) proposed a Bayesian approach, which modeled the spatially varying coefficient surface as a realization from a spatial process assuming a certain prior distribution of the coefficients. Another popular approach for estimating the SVCM in spatial regression analysis is through bivariate smoothing. For example, Fotheringham, Brunsdon, and Charlton (2002) proposed the geographically weighted regression (GWR) method, where the coefficient surface is locally estimated at each location via weighted least-squares with some distance-decay weights. Lu et al. (2009) extended the SVCM to adaptively varying-coefficient spatiotemporal models to deal with data that are observed irregularly over space and regularly in time. The SVCM has also been applied to FDA to deal with imaging response data. For example, in the image-on-scalar regression setting for Model (1), Zhu, Fan, and Kong (2014) proposed a local three-stage estimation method to fit the model using adaptive weights. However, their method relies heavily on estimating the spatial similarity and adaptive weights repeatedly, so it is computationally intensive for analyzing large imaging datasets. In addition, the local method also makes it difficult to conduct variable selection as we do in Section 3.

Another challenge in analyzing biomedical imaging data is the so-called "leakage" problem, which refers to the fact that the methods perform poorly when used to smooth data over complex domains. In many imaging studies, the objects (or regions) of interest on the images are usually irregularly shaped. Many conventional smoothing methods, such as kernel smoothing (Zhu, Fan, and Kong 2014), tensor product smoothing (Reiss and Ogden 2010), and wavelet smoothing (Morris and Carroll 2006), often encounter the "leakage" problem because they smooth inappropriately across boundary features (Ramsay 2002; Wang and Ranalli 2007). To overcome the "leakage" problem, we approximate the spatially varying coefficients in the functional linear regression model by using the bivariate splines over triangulation (BST; Lai and Schumaker 2007). Our method not only efficiently incorporates spatial information, but also preserves important features (shape and/or smoothness) of imaging data. Furthermore, our method is able to conveniently formulate a penalized function for sparse learning and model identification purposes, which cannot be easily implemented through other adaptive or sequential smoothing methods.

High-dimensional data occur very frequently in biomedical imaging studies. For example, in imaging genetics, it is of great interest to investigate how the effect of each genetic factor changes under the influence of multiple environmental variables, and in a typical gene-environment-wide association study (GEWAS), the number of SNPs could be extremely large, which poses a substantial challenge for applying SVCM and/or PLSVCM directly. When implementing the PLSVCM, another critical challenge lies in the identification of nonstationarity because users generally do not have prior knowledge on which coefficients are constant and which are varying over space. Hence, there is a strong interest in identifying important risk factors related to the imaging responses and in the meantime finding a more parsimonious semiparametric model structure that allows more efficient estimation. In this article, we first propose a BST-based estimator for the PLSVCM (with diverging p), and establish its consistency and asymptotic normality. We then propose a double-penalized approach for simultaneous sparse learning and model structure identification (i.e., determination of spatially varying vs. constant coefficients) in the presence of ultrahigh-dimensional covariates. In addition, we derive model selection consistency for the proposed method and show that it possesses the oracle property when the dimension of covariates exceeds the sample size.

We provide a new pathway to develop the methodology and formal theory for imaging-on-scalar regression with ultrahigh-dimensional predictors, in which the coefficient functions are approximated using the BST approach. Compared with the computationally expensive local smoothing methods (kernel, local polynomial), BST is a global smoothing method and allows easier implementation of piecewise polynomial representations of various degrees and various smoothness over an arbitrary triangulation. The BST approach was also used in Mu, Wang, and Wang (2018), Wang et al. (2020), and Yu et al. (2019). However, in these studies, either there was no covariate information or the dimension of the covariate vector was fixed. Thus, the results in these papers are not generally applicable to the SVCM or PLSVCM with a diverging number of covariates.

To the best of our knowledge, the present article is the first to develop theory and methodology for imaging-on-scalar regression in a high-dimensional setting. The SVCM and PLSVCM models we consider here are in a functional regression framework. The asymptotic analysis of such semiparametric models in high-dimensional settings is more complicated than those in the existing literature as we need to simultaneously deal with an infinite dimension of the functional response, spatial smoothing over a complex domain and ultrahigh-dimensional covariates. Another notorious difficulty in our theoretical investigation is that the covariates in the linear part and those in the spatially varying components could be dependent. To resolve the dependence between covariates in the constant and varying parts, we consider a projection of the covariates in the linear part onto the varying coefficient functional space, and study the properties of the projection onto the estimation space relative to the theoretical or empirical inner products introduced in Section 2.2; see Theorem 3.

The rest of the article is organized as follows. In Section 2, we consider the "oracle" PLSVCM, in which the active constant and varying index sets are treated as known. We adopt BST to approximate the unknown functional coefficients. The details of the estimation method for the "oracle" PLSVCM are given

in Section 2.1, and the asymptotic results of the proposed estimators are presented in Section 2.2. In Section 3, we describe the double-penalized bivariate spline procedure for selecting SVCMs when the dimension of covariates is high. Furthermore, we show the consistency of estimation, selection and model identification, and derive the asymptotic distribution of the proposed estimators. Section 4 demonstrates the performance of the proposed method through two simulation studies. In Section 5, we present our empirical analysis of the Alzheimer's Disease Neuroimaging Initiative (ADNI) data using the newly proposed procedure. Finally, we make some concluding remarks in Section 6. Proofs of the theorems, technical lemmas along with additional simulation studies are relegated to the supplementary materials.

2. PLSVCM and Its Estimation

Suppose the data consist of n individuals, and denote $\mathbf{X}_{(i)}$ as the covariates for the ith subject, that is, $\mathbf{X}_{(i)} = (X_{i1}, \ldots, X_{ip})^\top$. The observed measurements are $(Y_i(\boldsymbol{s}_j), \mathbf{X}_{(i)})$ for the ith individual, where $\{(Y_i(\boldsymbol{s}), \mathbf{X}_{(i)})\}_{i=1}^n$ are independent copies of $(Y(\boldsymbol{s}), \boldsymbol{X})$. For the scalar covariates X_{ik}'s, let $\mathcal{A}_c$ and $\mathcal{A}_v$ denote the index sets of finite covariates with constant and varying coefficients, respectively. We first study the estimation for the following partially linear spatially varying coefficient model (PLSVCM) by assuming that $\mathcal{A}_c$ and $\mathcal{A}_v$ are known sets with finite cardinalities:

$$Y_i(\boldsymbol{s}_j) = \sum_{k\in\mathcal{A}_c} X_{ik}\alpha_k + \sum_{k\in\mathcal{A}_v} X_{ik}\beta_k(\boldsymbol{s}_j) + \eta_i(\boldsymbol{s}_j) + \varepsilon_{ij}, \quad \boldsymbol{s}_j \in \Omega, \quad (2)$$

for $i = 1, \ldots, n$, $j = 1, \ldots, N_s$, where ε_{ij}'s are independent random errors. The analyses for the SVCM Model (1) with ultrahigh-dimension p will be presented in Section 3. For model identifiability, we assume that $\int_\Omega \beta_k(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = 0$ to ensure that there is no constant effect in the nonlinear function $\beta_k(\boldsymbol{s})$.

2.1. The Estimation Method

Our estimation is based on basis approximations using bivariate splines over triangulations. The idea is to approximate the function $\beta_k(\cdot)$ by splines which are piecewise polynomial bivariate functions over a two-dimensional triangulated domain. We use this approximation to construct least squares estimators of the linear and nonlinear components of the model. In the following subsections, we describe the background of triangulations and introduce the spline estimators.

2.1.1. Triangulations

Triangulation is an effective strategy to handle data distribution on irregular regions with complex boundaries and/or interior holes. It has attracted substantial recent attention in many applied areas, such as geo-spatial studies, numerical solutions of partial differential equations, image enhancements, and computer aided geometric design; see, for example, the recent comprehensive book of Lai and Schumaker (2007).

We use τ to denote a triangle which is a convex hull of three points not located in one line. A collection $\triangle = \{\tau_1, \ldots, \tau_N\}$ of N triangles is called a triangulation of $\Omega = \cup_{j=1}^N \tau_j$ provided that the intersection of any pair of triangles in $\triangle$ is either empty, a common vertex, or a common edge. In general, any kind of polygon shape can be used for the partition of Ω. In this article, we restrict our attention to triangulations of Ω because any polygonal domain of arbitrary shape can be partitioned into finitely many triangles. In the following, we assume that all $\boldsymbol{s}_j$'s are inside triangles of $\triangle$.

Given a triangle $\tau \in \triangle$, let $|\tau|$ be its longest edge length, and let R_τ be the radius of the largest disk which can be inscribed in τ. Define the shape parameter of τ as the ratio $\beta_\tau = |\tau|/R_\tau$. When β_τ is small, the triangles are relatively uniform in the sense that all angles of triangles in the triangulation τ are relatively similar. Denote the size of $\triangle$ by $|\triangle| := \max\{|\tau|, \tau \in \triangle\}$, that is, the length of the longest edge of $\triangle$. In the rest of the article, we say a triangulation $\triangle$ is π-quasi-uniform if there is a positive π such that $\triangle$ satisfies for all $\tau \in \triangle$, $|\triangle|/R_\tau \le \pi$.

2.1.2. Bivariate Spline Estimators

For a nonnegative integer r, let $\mathbb{C}^r(\Omega)$ be the collection of all rth continuously differentiable functions over Ω. Given a triangulation $\triangle$, let $\mathcal{S}_d^r(\triangle) = \{s \in \mathbb{C}^r(\Omega) : s|_\tau \in \mathcal{P}_d(\tau), \tau \in \triangle\}$ be a spline space of degree d and smoothness r over the triangulation $\triangle$, where $s|_\tau$ is the polynomial piece of spline s restricted on the triangle τ, and $\mathcal{P}_d$ is the space of all polynomials of degree less than or equal to d.

For any $k = 1, \ldots, p$, let $\triangle_k$ be the triangulation for the kth component, and let $\{B_{k\ell}\}_{\ell\in\mathcal{J}_k}$ be a set of bivariate Bernstein basis polynomials for $\mathcal{S}_d^r(\triangle_k)$, where $\mathcal{J}_k$ is the index set of all the spline basis functions for the kth component. See Section S.1 of the supplementary materials for a more detailed introduction of the Bernstein basis polynomials. Then, we can write the function $\beta_k(\boldsymbol{s}) \approx \sum_{\ell\in\mathcal{J}_k} B_{k\ell}(\boldsymbol{s})c_{k\ell} = B_k^\top(\boldsymbol{s})\boldsymbol{c}_k$, where $\boldsymbol{c}_k = (c_{k\ell}, \ell \in \mathcal{J}_k)^\top$ is the spline coefficient vector, and $\mathbf{B}_k(\cdot) = (B_{k\ell}(\cdot), \ell \in \mathcal{J}_k)^\top$ is the vector of bivariate basis functions. To meet the smoothness requirement of the splines, we need to impose some constraints on the spline coefficients. The smoothness conditions are linear. Let $\mathbf{H}_k$ denote the constraint matrix on the coefficients $\boldsymbol{c}_k$, which depends on the smoothness parameter r_k and the structure of the triangulation and enforces smoothness across shared edges of triangles; see Yu et al. (2019). Put all smoothness conditions together to write $\mathbf{H}_k\boldsymbol{c}_k = \mathbf{0}$. Without loss of generality, we assume $\mathbf{B}_1(\boldsymbol{s}) = \cdots = \mathbf{B}_p(\boldsymbol{s})$, and denote it as $\mathbf{B}(\boldsymbol{s}) = (B_\ell(\boldsymbol{s}), \ell \in \mathcal{J})^\top$. Similarly, we assume $\mathbf{H} = \mathbf{H}_1 = \cdots = \mathbf{H}_p$. For practitioners, the construction of bivariate spline basis $\mathbf{B}$ and the constraint matrix $\mathbf{H}$ can be easily done via our new R package `Basis`. Then we can approximate the nonparametric function $\beta_k(\boldsymbol{s})$, $k = 1, \ldots, p$, using the normalized triangulation splines as

$$\beta_k(\boldsymbol{s}) \approx \beta_{nk}(\boldsymbol{s}) = \mathbf{B}^\top(\boldsymbol{s})\boldsymbol{c}_k. \quad (3)$$

For the ith subject, we observe a real-valued imaging measure $Y_i(\boldsymbol{s}_j)$ at point $\boldsymbol{s}_j \in \Omega$ for $j = 1, \ldots, N_s$. For example, $\boldsymbol{s}_j = (s_{j1}, s_{j2})^\top$ could be the coordinates of a voxel center in a two-dimensional cross-section of a three-dimensional brain image.

Given $\{(\mathbf{X}_{(i)}, Y_i(\boldsymbol{s}_j)) : i = 1, \ldots, n, j = 1, \ldots, N_s\}$, we propose to minimize

$$\sum_{i=1}^n \sum_{j=1}^{N_s} \left\{ Y_i(\boldsymbol{s}_j) - \sum_{k\in\mathcal{A}_c} X_{ik}\alpha_k - \sum_{k\in\mathcal{A}_v} X_{ik}\beta_k(\boldsymbol{s}_j) \right\}^2. \quad (4)$$

We now approximate the varying coefficient functions by using the bivariate basis functions introduced. Let $Y_{ij} = Y_i(\boldsymbol{s}_j)$, then combining (3) and (4), we aim to minimize the following objective function

$$L_n(\boldsymbol{\alpha}, \boldsymbol{c}) = \sum_{i=1}^{n}\sum_{j=1}^{N_s}\left\{Y_{ij} - \sum_{k\in\mathcal{A}_c} X_{ik}\alpha_k - \sum_{k\in\mathcal{A}_v} X_{ik}\mathbf{B}^\top(\boldsymbol{s}_j)\boldsymbol{c}_k\right\}^2, \tag{5}$$

subject to $\mathbf{H}\boldsymbol{c}_k = \mathbf{0}$. We remove the constraint via QR decomposition of $\mathbf{H}^\top$. By the QR decomposition, we have $\mathbf{H}^\top = \mathbf{QR} = (\mathbf{Q}_1\ \mathbf{Q}_2)\binom{\mathbf{R}_1}{\mathbf{R}_2}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{R}$ is an upper triangle matrix, the submatrix $\mathbf{Q}_1$ is the first r columns of $\mathbf{Q}$, with r being the rank of $\mathbf{H}$, and $\mathbf{R}_2$ is a matrix of zeros. Then, the constrained optimization problem in (5) is reduced to the following non-constrained optimization problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} L_n(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=1}^{N_s}\left[Y_{ij} - \sum_{k\in\mathcal{A}_c} X_{ik}\alpha_k - \sum_{k\in\mathcal{A}_v} X_{ik}\{\mathbf{B}^*(\boldsymbol{s}_j)\}^\top\boldsymbol{\gamma}_k\right]^2,$$

where $\mathbf{B}^*(\boldsymbol{s}) = \mathbf{Q}_2^\top\mathbf{B}(\boldsymbol{s})$.

To ensure $\int_\Omega \beta_k(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = 0$, we consider the normalized bivariate spline basis of $\mathbf{B}^*(\boldsymbol{s}) = \{B_0^*(\boldsymbol{s}), \ldots, B_{J_n}^*(\boldsymbol{s})\}$. For any $\ell \geq 1$, define the centered bivariate spline basis as $B_\ell^0(\cdot) = B_\ell^*(\cdot) - B_0^*(\cdot)\int_\Omega B_\ell^*(\boldsymbol{s})\mathrm{d}\boldsymbol{s}/\int_\Omega B_0^*(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$, and rescale it by $B_\ell^N(\cdot) = B_\ell^0(\cdot)/\|B_\ell^0(\cdot)\|_{L^2(\Omega)}$, $\ell = 1, \ldots, J_n$, where $\|f\|_{L^2(\Omega)}^2 = \int_\Omega f^2(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$ for any function f. At the risk of abusing the notation, below we still use $B_\ell(\cdot)$ instead of $B_\ell^N(\cdot)$ to denote the normalized spline basis to avoid creating too many new symbols. Denote $\mathbf{B}(\boldsymbol{s}) = \{B_\ell(\boldsymbol{s}), \ell = 1, \ldots, J_n\}$ the collection of all the normalized basis functions. We choose to use the normalized bivariate spline basis functions to simplify the proof. The results in the following sections still hold if centered but nonnormalized basis functions are used.

Thus, one obtains the estimators $\widehat{\alpha}_k^o$, $k \in \mathcal{A}_c$ and $\widehat{\boldsymbol{\gamma}}_k^o$, $k \in \mathcal{A}_v$, by minimizing the following non-constrained objective function with normalized bivariate spline basis

$$L_n(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=1}^{N_s}\left\{Y_{ij} - \sum_{k\in\mathcal{A}_c} X_{ik}\alpha_k - \sum_{k\in\mathcal{A}_v} X_{ik}\mathbf{B}^\top(\boldsymbol{s}_j)\boldsymbol{\gamma}_k\right\}^2. \tag{6}$$

Consequently, the estimator of $\beta_k(\cdot)$ is $\widehat{\beta}_k^o(\boldsymbol{s}) = \mathbf{B}^\top(\boldsymbol{s})\widehat{\boldsymbol{\gamma}}_k^o$, $k \in \mathcal{A}_v$.

2.2. Theoretical Properties

This section studies the asymptotic properties of the proposed estimators. Throughout, we will use an index 0 to denote the true parameter values and functions in model (2). To discuss these properties, we introduce some notation of norms. For any function g over the closure of domain Ω, denote $\|g\|_{L^2(\Omega)}^2 = \int_{\boldsymbol{s}\in\Omega} g^2(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$ the regular L_2 norm of g, and $\|g\|_{\infty,\Omega} = \sup_{\boldsymbol{s}\in\Omega}|g(\boldsymbol{s})|$ the supremum norm of g. For directions s_j, $j = 1, 2$, let $D_{s_j}^q g(\boldsymbol{s})$ denote the qth order derivative in the direction s_j at the point $\boldsymbol{s}$. Let $|g|_{\upsilon,\infty,\Omega} = \max_{i+j=\upsilon}\|D_{s_1}^i D_{s_2}^j g(\boldsymbol{s})\|_{\infty,\Omega}$ be the maximum norms of all the υth order derivatives of g over Ω. Let $\mathcal{W}^{\ell,\infty}(\Omega) = \left\{g : |g|_{k,\infty,\Omega} < \infty, 0 \leq k \leq \ell\right\}$ be the standard Sobolev space. To avoid confusion, let $(\alpha_{0k}, k \in \mathcal{A}_c)$ and $(\beta_{0k}, k \in \mathcal{A}_v)$ be the true parameter values below.

In the following, we first introduce some technical assumptions.

Assumption 1 (Model). For any $k \in \mathcal{A}_v$, the varying coefficient function $\beta_{0k} \in \mathcal{W}^{d+1,\infty}(\Omega)$ for an integer $d \geq 0$.

Assumption 2 (Covariates). For any $k = 1, \ldots, p$, there exists a positive constant C_X such that $\mathrm{E}|X_k|^6 \leq C_X$. The eigenvalues of $\boldsymbol{\Sigma}_X = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top)$ are bounded away from 0 and infinity, where $\boldsymbol{X}$ is a generic p-dimensional vector of all covariates.

Assumption 3 (Errors). For any $i = 1, \ldots, n$, $j = 1, \ldots, N_s$, ε_{ij}'s are independent with mean 0 and variance σ_ε^2; the variance function for $\eta_i(\boldsymbol{s})$, $G_\eta(\boldsymbol{s}, \boldsymbol{s})$, satisfies $0 < c_G \leq G_\eta(\boldsymbol{s}, \boldsymbol{s}) \leq C_G \leq \infty$ for any $\boldsymbol{s} \in \Omega$.

Assumption 4 (Resolution of images). If the location points $\boldsymbol{s}$ are deterministic, we assume that $\sup_{\boldsymbol{s}\in\Omega}|Q_{N_s}(\boldsymbol{s}) - Q(\boldsymbol{s})| = O(N_s^{-1/2})$, where $Q_{N_s}(\boldsymbol{s}) = N_s^{-1}\sum_{j=1}^{N_s} I(s_{j1} \leq s_1, s_{j2} \leq s_2)$ is the empirical cumulative distribution function and $Q(\boldsymbol{s})$ is a distribution with a positive continuous density.

Assumption 5 (Triangulations). The triangulation $\triangle$ is π-quasi-uniform.

Assumption 6 (Size of triangulation). As $N_s \to \infty$, $n \to \infty$, $N_s|\triangle|^2 \to \infty$ and $|\triangle|^{2(d+1)}n \to 0$.

Remark 1. Assumptions 1–3 are regularity conditions that are commonly used in the high-dimension literature. Assumptions 4 and 5 are regular conditions that are widely used in the triangulation bases literature. Assumption 6 describes the requirement on the sample size n, the number of voxels N_s on each individual image and the triangulation size relative to n and N_s. Note that the size of the triangulation $|\triangle|$ plays the role of smoothing parameter. Assumption 6 requires that $N_s^{-1/2} \ll |\triangle| \ll n^{-1/\{2(d+1)\}}$, which is a wide range given that the number of pixels are typically much larger than the sample size. It also suggests that the precise determination of the size of triangulations is not a particular concern in terms of asymptotic behavior. In Section 4.1, we give some discussion and advice on how to choose the triangulation.

The following theorem establishes the consistency of the proposed constant estimators and spline estimators for the PLSVCM.

Theorem 1. Suppose that Assumptions 1–6 and model (2) hold. Then

(i) $\sum_{k\in\mathcal{A}_c}(\widehat{\alpha}_k^o - \alpha_{0k})^2 = O_P\left\{n^{-1} + n^{-1}N_s^{-1}|\triangle|^{-2} + |\triangle|^{2(d+1)}\right\}$,

(ii) $\sum_{k\in\mathcal{A}_v}\|\widehat{\beta}_k^o - \beta_{0k}\|_{L_2(\Omega)}^2 = O_P\left\{n^{-1} + n^{-1}N_s^{-1}|\triangle|^{-2} + |\triangle|^{2(d+1)}\right\}$.

Remark 2. Result (i) shows the rate of convergence of $\widehat{\alpha}_k^o$. Result (ii) provides the global convergence rate of $\widehat{\beta}_k^o$. The bias of $\widehat{\alpha}_k^o$ and $\widehat{\beta}_k^o$ is of order $O_P(|\triangle|^{(d+1)})$ and the noise is of the order $O_P(n^{-1/2} + n^{-1/2}N_s^{-1/2}|\triangle|)$. Note that under Assumption 6, $N_s|\triangle|^2 \to \infty$ and $|\triangle|^{2(d+1)}n \to 0$, as $N_s \to \infty$ and $n \to \infty$. Thus, $\widehat{\alpha}_k^o$ is root-n consistent. Finally, the convergence rate we obtained for the bivariate spline estimator $\widehat{\beta}_k^o$ is similar to the one for the estimated univariate coefficient function in the varying coefficient model for dense longitudinal observations in Huang, Wu, and Zhou (2004).

Next we establish the asymptotic normal distribution for the estimators of the constant coefficients. To overcome the obstacle that covariates in the linear part and those in the spatially varying components could be dependent, we consider a projection of covariates in the linear part onto the varying coefficient functional space, and study the properties of the projection onto the estimation space relative to the theoretical or empirical inner product introduced. To be specific, for any index set $\mathcal{I}$, denote $\boldsymbol{X}_{\mathcal{I}} = (\boldsymbol{X}_k, k \in \mathcal{I})^\top$, and for any vector $\boldsymbol{a}$, let $\boldsymbol{a}_{\mathcal{I}} = (a_k, k \in \mathcal{I})^\top$. To facilitate our discussion, we introduce the following two inner products. Let $\boldsymbol{x} = (\boldsymbol{x}_{\mathcal{A}_c}^\top, \boldsymbol{x}_{\mathcal{A}_v}^\top)^\top$. Similarly, we write $\boldsymbol{X}$ in a parallel fashion. For $g^{(1)}(\boldsymbol{x}, \boldsymbol{s})$ and $g^{(2)}(\boldsymbol{x}, \boldsymbol{s})$, define the empirical inner product as

$$\langle g^{(1)}, g^{(2)}\rangle_{n,N_s} = \frac{1}{nN_s}\sum_{i=1}^{n}\sum_{j=1}^{N_s} g^{(1)}(\mathbf{X}_{(i)}, \boldsymbol{s}_j)g^{(2)}(\mathbf{X}_{(i)}, \boldsymbol{s}_j), \quad (7)$$

and the theoretical inner product as

$$\langle g^{(1)}, g^{(2)}\rangle = \mathrm{E}\left\{\int_\Omega g^{(1)}(\boldsymbol{X}, \boldsymbol{s})g^{(2)}(\boldsymbol{X}, \boldsymbol{s})\mathrm{d}Q(\boldsymbol{s})\right\}, \quad (8)$$

and denote the corresponding empirical and theoretical norms as $\|\cdot\|_{n,N_s}$ and $\|\cdot\|$. For any $k = 1, \ldots, p$, let $x_k(\boldsymbol{x}, \boldsymbol{s}) = x_k$ be a function of $(\boldsymbol{x}, \boldsymbol{s})$ that maps to the kth element of $\boldsymbol{x}$. Define $\mathcal{F}_+ = \{F(\boldsymbol{x}, \boldsymbol{s}) = \sum_{k\in\mathcal{A}_v} x_k g_k(\boldsymbol{s}) : \int_\Omega g_k(\boldsymbol{s})\mathrm{d}Q(\boldsymbol{s}) = 0\}$. For any $k \in \mathcal{A}_c$, let

$$\begin{aligned}\Gamma_k(\cdot,\cdot) &= \underset{F(\cdot,\cdot)\in\mathcal{F}_+}{\arg\min}\ \mathrm{E}\left[\int_\Omega \{X_{ik} - F(\mathbf{X}_{(i)}, \boldsymbol{s})\}^2\mathrm{d}Q(\boldsymbol{s})\right] \\ &= \underset{F(\cdot,\cdot)\in\mathcal{F}_+}{\arg\min}\ \|x_k - F\|^2 \end{aligned} \quad (9)$$

be the orthogonal projection of x_k onto $\mathcal{F}_+$ relative to the theoretical inner product in (8). To be more specific, we can express $\Gamma_k(\boldsymbol{x}, \boldsymbol{s})$ as

$$\Gamma_k(\boldsymbol{x}, \boldsymbol{s}) = \sum_{k'\in\mathcal{A}_v} x_{k'} g_{k,k'}^o(\boldsymbol{s}), \quad (10)$$

for some $g_{k,k'}^o(\boldsymbol{s})$ satisfying $\int_\Omega g_{k,k'}^o(\boldsymbol{s})\mathrm{d}Q(\boldsymbol{s}) = 0$, for any $k \in \mathcal{A}_c$.

Let $\boldsymbol{\Gamma}_{\mathcal{A}_c}(\boldsymbol{X}, \boldsymbol{s}) = \{\Gamma_k(\boldsymbol{X}, \boldsymbol{s}), k \in \mathcal{A}_c\}^\top$ and denote $\Sigma_e(\boldsymbol{s}, \boldsymbol{s}') = G_\eta(\boldsymbol{s}, \boldsymbol{s}') + \sigma^2(\boldsymbol{s})I(\boldsymbol{s} = \boldsymbol{s}')$. Denote two matrices:

$$\boldsymbol{\Xi} = \mathrm{E}\int_\Omega \{\boldsymbol{X}_{\mathcal{A}_c} - \boldsymbol{\Gamma}_{\mathcal{A}_c}(\boldsymbol{X}, \boldsymbol{s})\}\{\boldsymbol{X}_{\mathcal{A}_c} - \boldsymbol{\Gamma}_{\mathcal{A}_c}(\boldsymbol{X}, \boldsymbol{s})\}^\top\mathrm{d}\boldsymbol{s}, \quad (11)$$

$$\begin{aligned}\boldsymbol{\Xi}_e = \mathrm{E}\int_{\Omega^{\otimes 2}} &\{\boldsymbol{X}_{\mathcal{A}_c} - \boldsymbol{\Gamma}_{\mathcal{A}_c}(\boldsymbol{X}, \boldsymbol{s})\}\Sigma_e(\boldsymbol{s}, \boldsymbol{s}') \\ &\times \{\boldsymbol{X}_{\mathcal{A}_c} - \boldsymbol{\Gamma}_{\mathcal{A}_c}(\boldsymbol{X}, \boldsymbol{s}')\}^\top\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{s}'.\end{aligned}$$

To make $\boldsymbol{\alpha}_{0,\mathcal{A}_c}$ estimable at the $\sqrt{n}$ rate, we need a condition to ensure that $\boldsymbol{X}_{\mathcal{A}_c}$ and $\boldsymbol{X}_{\mathcal{A}_v} \otimes \boldsymbol{B}$ are not functionally related. In addition, we need to assume that, for any $k \in A_c, k' \in A_v$, the coefficient function $g_{k,k'}^o(\cdot)$ defined in (10) can be approximated by functions in the bivariate spline space. So next, we introduce the following two assumptions:

Assumption 7 (Covariance matrix). The matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Xi}_e$ are positive definite.

Assumption 8 (Coefficient functions). The coefficients functions in (10), $g_{k,k'}^o(\cdot) \in \mathcal{W}^{d+1,\infty}(\Omega)$, $k \in A_c, k' \in \mathcal{A}_v$.

Theorem 2. Suppose Assumptions 1–8 and model (2) hold. If for any $k \in \mathcal{A}_c$, $|X_{ik}| \le C_k < \infty$, then $(\mathbf{V}_n^c)^{-1/2}\left(\widehat{\boldsymbol{\alpha}}^o - \boldsymbol{\alpha}_{0,\mathcal{A}_c}\right) \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_{|\mathcal{A}_c|})$, as $N_s \to \infty$ and $n \to \infty$, where $\mathbf{I}_{|\mathcal{A}_c|}$ is an $|\mathcal{A}_c| \times |\mathcal{A}_c|$ identity matrix, and

$$\mathbf{V}_n^c = \Sigma_c^{-1}\Sigma_{c,e}\Sigma_c^{-1}, \quad (12)$$

with $\Sigma_c = (nN_s)^{-1}\mathbf{D}_c^\top\mathbf{D}_c$, and $\Sigma_{c,e} = (nN_s)^{-2}\mathbf{D}_c^\top\mathrm{diag}\{\Sigma_{i,e}\}_{i=1}^n\mathbf{D}_c$ for $\mathbf{D}_c$ given in (S.16) and $\Sigma_{i,e} \equiv \Sigma_e = \{\Sigma_e(\boldsymbol{s}_j, \boldsymbol{s}_{j'})\}_{j,j'=1}^{N_s} = \{G_\eta(\boldsymbol{s}_j, \boldsymbol{s}_{j'})\}_{j,j'=1}^{N_s} + \mathrm{diag}\{\sigma^2(\boldsymbol{s}_j), j = 1, \ldots, N_s\}$.

Remark 3. For any $i = 1, \ldots, n$, let $\widehat{\boldsymbol{\mu}}_i = \{\widehat{\mu}_i(\boldsymbol{s}_j)\}_{j=1}^{N_s}$, where $\widehat{\mu}_i(\boldsymbol{s}_j) = \sum_{k\in\mathcal{A}_c} X_{ik}\widehat{\alpha}_k^o + \sum_{k\in\mathcal{A}_v} X_{ik}\widehat{\beta}_k^o(\boldsymbol{s}_j)$. Note that by Lemma S.15 in the supplementary materials, $\Sigma_{i,e}$ can be consistently estimated by $\widehat{\Sigma}_e = (n - |\mathcal{A}_c|)^{-1}\sum_{i=1}^n\{\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i\}\{\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i\}^\top$. Consequently, $\mathbf{V}_n^c$ can be estimated by $\widehat{\mathbf{V}}_n^c = \Sigma_c^{-1}\widehat{\Sigma}_{c,e}\Sigma_c^{-1}$, where $\widehat{\Sigma}_{c,e} = (nN_s)^{-2}\mathbf{D}_c^\top\mathrm{diag}\{\widehat{\Sigma}_{i,e}\}_{i=1}^n\mathbf{D}_c$.

Theorem 2 provides the sample variance-covariance matrix for the estimator $\widehat{\boldsymbol{\alpha}}^o$. To derive the form of the asymptotic covariance matrix of $\widehat{\boldsymbol{\alpha}}^o$, we define the space $\mathcal{F}_{n,+} = \{F(\boldsymbol{x}, \boldsymbol{s}) = \sum_{k\in\mathcal{A}_v} x_k g_k(\boldsymbol{s}) : g_k(\boldsymbol{s}) \in \mathcal{S}_d^r(\triangle) \cap \mathcal{H}^2\}$, where $\mathcal{S}_d^r(\triangle)$ is spline space for with degree d and smoothness r over triangulation $\triangle$, as defined in Section 2.1.2, and $\mathcal{H}^2$ is space for normalized functions. For any $k \in \mathcal{A}_c$, in a parallel fashion to (9), we define

$$\begin{aligned}\widehat{\Gamma}_{n,k}(\cdot,\cdot) &= \underset{F(\cdot,\cdot)\in\mathcal{F}_{n,+}}{\arg\min}\ \frac{1}{nN_s}\sum_{i=1}^{n}\sum_{j=1}^{N_s}\{X_{ik} - F(\mathbf{X}_{(i)}, \boldsymbol{s}_j)\}^2 \\ &= \underset{F(\cdot,\cdot)\in\mathcal{F}_{n,+}}{\arg\min}\ \|x_k - F\|_{n,N_s}^2 \end{aligned} \quad (13)$$

as the orthogonal projection of x_k onto $\mathcal{F}_{n,+}$ relative to the empirical inner product given in (7). Let $\widehat{\Pi}_n$ and Π_n denote the projection operator onto $\mathcal{F}_{n,+}$ relative to the norm $\|\cdot\|_{n,N_s}$ and $\|\cdot\|$, respectively.

The next result shows that $\widehat{\Gamma}_{n,k}$ given in (13) is a consistent estimator of Γ_k defined in (9).

Theorem 3. Suppose Assumptions 1–8 and model (2) hold. For any $k \in \mathcal{A}_c$, we have $\|\widehat{\Gamma}_{n,k} - \Gamma_k\|_{n,N_s}^2 = o_P(1)$.

Remark 4. Theorem 3 can be used to derive the asymptotic form of variance-covariance matrix $\mathbf{V}_n^c$ defined in (12). As shown in Section S.4 of the supplementary materials, for $k, k' \in \mathcal{A}_c$, $\boldsymbol{\Sigma}_c$ has the (k, k')th entry $\langle x_k - \widehat{\Gamma}_{n,k}(x_k, \boldsymbol{s}), x_{k'} - \widehat{\Gamma}_{n,k'}(x_{k'}, \boldsymbol{s})\rangle_{n,N_s}$, which is the empirical inner product of the projection residuals

of x_k, $x_{k'}$ onto the space $\mathcal{F}_{n,+}$. By Theorem 3, it converges to $\langle x_k - \Gamma_k(x_k, \boldsymbol{s}), x_{k'} - \Gamma_{k'}(x_{k'}, \boldsymbol{s})\rangle$, which is the (k, k')th entry of $\boldsymbol{\Xi}$, in probability. Similarly, it can be shown that the (k, k')th entry of $\boldsymbol{\Sigma}_{\mathrm{c,e}}$ converges to the (k, k')th entry of $\boldsymbol{\Xi}_{\mathrm{e}}$ in probability.

The following corollary provides the asymptotic variance-covariance matrix of the estimator $\widehat{\boldsymbol{\alpha}}^o$, which is a direct result of Theorems 2 and 3.

Corollary 1. Suppose Assumptions 1–8 and model (2) hold. As $N_s \to \infty$ and $n \to \infty$, $n^{1/2}\left(\widehat{\boldsymbol{\alpha}}^o - \boldsymbol{\alpha}_{0,\mathcal{A}_{\mathrm{c}}}\right) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Xi}^{-1}\boldsymbol{\Xi}_{\mathrm{e}}\boldsymbol{\Xi}^{-1})$, where $\boldsymbol{\Xi}$ and $\boldsymbol{\Xi}_{\mathrm{e}}$ are given in (11).

3. Sparse Model Identification and Learning Estimator

In this section, we consider the SVCM in (1), and propose a sparse model identification and learning estimation (referred to as SMILE) procedure to simultaneously identify and estimate the nonzero spatially varying and constant coefficients. There have been several recent attempts to solve this problem for VCM (see Noh, Chung, and Van Keilegom 2012; Wang and Kulasekera 2012; Lian, Lai, and Liang 2013; Li, Ke, and Zhang 2015; Lian, Meng, and Zhao 2015; Chen, Bai, and Fung 2017). However, all of them studied the VCM with time index or other univariate indices. In contrast, our methodology is developed under the SVCM framework with a bivariate index, which requires more advanced tools to deal with the spatial index and corresponding irregular domain. In addition, the works in the existing literature can only be viewed as a case with $n = 1$ and $N \to \infty$, and they cannot handle functional responses. In our framework, both N and n go to infinity, which poses challenges in both theory and practice.

We evaluate the proposed selection and estimation method in the high-dimensional setting, that is, we allow the dimension p to be much larger than the sample size n. To decide if $f_k(\boldsymbol{s})$ is varying, for each $k = 1, \ldots, p$, we can decompose f_k into a constant and a nonlinear part: $f_k(\boldsymbol{s}) = \alpha_k + \beta_k(\boldsymbol{s})$, where $\beta_k(\boldsymbol{s})$ is some unknown smooth nonlinear function. For model identifiability, we assume that $\int_\Omega \beta_k(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = 0$, and this constraint ensures there is no constant effect in nonlinear function $\beta_k(\boldsymbol{s})$. Then, we can express model (1) as

$$Y_i(\boldsymbol{s}) = \sum_{k=1}^{p} X_{ik}\alpha_k + \sum_{k=1}^{p} X_{ik}\beta_k(\boldsymbol{s}) + \eta_i(\boldsymbol{s}) + \varepsilon_i(\boldsymbol{s}),$$
$$i = 1, \ldots, n,\ \boldsymbol{s} \in \Omega. \tag{14}$$

In the following, we say X_k has a constant effect on the response if $\alpha_k \neq 0$ and $\beta_k(\boldsymbol{s}) = 0$ for all $\boldsymbol{s} \in \Omega$, and X_k has a varying effect on the response if $\beta_k(\boldsymbol{s})$ is not zero for some $\boldsymbol{s}$. Recall that our interest lies in selecting variables with nonzero varying and constant effects. In practice, some of the covariates may be irrelevant to the response variable, with the corresponding varying-coefficient functions being zero almost surely. So we identify the irrelevant covariates and estimate the nonzero coefficient functions for the relevant ones simultaneously. Explicitly, we define the following index sets for $\boldsymbol{X}$:

$$\begin{aligned}
\text{Active constant index set for } \boldsymbol{X}: &\mathcal{A}_{\mathrm{c}} = \{k = 1, \ldots, \\
&\quad p : \alpha_k \neq 0, \beta_k(\cdot) \equiv 0\}, \\
\text{Active varying index set for } \boldsymbol{X}: &\mathcal{A}_{\mathrm{v}} = \{k = 1, \ldots, \\
&\quad p : \beta_k(\cdot) \neq 0\}, \\
\text{Inactive index set for } \boldsymbol{X}: &\mathcal{N} = \{k = 1, \ldots, \\
&\quad p : \alpha_k \equiv 0, \beta_k(\cdot) \equiv 0\}.
\end{aligned} \tag{15}$$

Accordingly, we define the active index set for $\boldsymbol{X}$ as $\mathcal{A} = \mathcal{A}_{\mathrm{c}} \cup \mathcal{A}_{\mathrm{v}}$.

3.1. Adaptive Group Lasso Estimator

The model selection problem for model (14) is equivalent to the problem of identifying $\mathcal{A}_{\mathrm{c}}$ and $\mathcal{A}_{\mathrm{v}}$. To achieve this, given $\{(\mathbf{X}_{(i)}, Y_i(\boldsymbol{s}_j)) : i = 1, \ldots, n, j = 1, \ldots, N_s\}$, we propose to minimize

$$\begin{aligned}
&L_n(\boldsymbol{\alpha}, \boldsymbol{\gamma};\ \rho_{n1}, \rho_{n2}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{N_s}\left\{Y_i(\boldsymbol{s}_j) - \sum_{k=1}^{p} X_{ik}\alpha_k - \sum_{k=1}^{p} X_{ik}\mathbf{B}^\top(\boldsymbol{s}_j)\boldsymbol{\gamma}_k\right\}^2 \\
&+ \sum_{k=1}^{p} p_{\rho_{n1}}(|\alpha_k|) + \sum_{k=1}^{p} p_{\rho_{n2}}(\|\boldsymbol{\gamma}_k\|),
\end{aligned} \tag{16}$$

where $\|\boldsymbol{\gamma}_k\|$ is the Euclidean norm of $\boldsymbol{\gamma}_k$, and $p_{\rho_{n1}}(\cdot)$ and $p_{\rho_{n2}}(\cdot)$ are penalty functions chosen as adaptive lasso penalty functions. To be specific, we choose $p_{\rho_{n1}}(|\alpha_k|) = \rho_{n1} w^{\mathrm{c}}_{n,k}|\alpha_k|$, and $p_{\rho_{n2}}(\|\boldsymbol{\gamma}_k\|) = \rho_{n2} w^{\mathrm{v}}_{n,k}\|\boldsymbol{\gamma}_k\|$, where $w^{\mathrm{c}}_{n,k} = |\widetilde{\alpha}_k|^{-1}$ and $w^{\mathrm{v}}_{n,k} = \|\widetilde{\boldsymbol{\gamma}}_k\|^{-1}$, with $\widetilde{\alpha}_k$ and $\widetilde{\boldsymbol{\gamma}}_k$ being some consistent initial estimators for α_k and $\boldsymbol{\gamma}_k$, respectively; ρ_{n1} and ρ_{n2} are regularization parameters controlling the amount of shrinkage, and ρ_{n1} and $\rho_{n2} \to \infty$ as $N_s \to \infty$ and $n \to \infty$. Details regarding the choice of initial estimators $\widetilde{\alpha}_k$ and $\widetilde{\boldsymbol{\gamma}}_k$ can be found in the discussion about Assumption 12 in Section 3.2.

By minimizing (16), we obtain the estimator $\widehat{\alpha}_k$ and $\widehat{\boldsymbol{\gamma}}_k$, and consequently, the estimator of $\beta_k(\cdot)$ is $\widehat{\beta}_k(\boldsymbol{s}) = \widehat{\boldsymbol{\gamma}}_k^\top \mathbf{B}(\boldsymbol{s})$, $k = 1, \ldots, p$. Therefore, we obtain the estimator of $f_k(\cdot)$ as follows: $\widehat{f}_k(\boldsymbol{s}) = \widehat{\alpha}_k + \mathbf{B}^\top(\boldsymbol{s})\widehat{\boldsymbol{\gamma}}_k$, $k = 1, \ldots, p$.

Then, the model structure selected is defined by

$$\begin{aligned}
\widehat{\mathcal{A}}_{\mathrm{c}} &= \left\{k : |\widehat{\alpha}_k| \neq 0, \|\widehat{\boldsymbol{\gamma}}_k\| = 0, 1 \le k \le p\right\}, \\
\widehat{\mathcal{A}}_{\mathrm{v}} &= \left\{k : \|\widehat{\boldsymbol{\gamma}}_k\| \neq 0, 1 \le k \le p\right\}, \\
\widehat{\mathcal{N}} &= \left\{k : |\widehat{\alpha}_k| = 0, \|\widehat{\boldsymbol{\gamma}}_k\| = 0, 1 \le k \le p\right\}.
\end{aligned} \tag{17}$$

3.2. Technical Assumptions

This section studies the asymptotic properties of the proposed penalized estimators. In the following, we first introduce some technical assumptions.

Assumption 9 (Model). The numbers of nonzero components $|\mathcal{A}_{\mathrm{c}}|$ and $|\mathcal{A}_{\mathrm{v}}|$ are fixed; there exist constants $c_\alpha > 0$ and $c_\beta > 0$ such that $\min_{k\in\mathcal{A}_{\mathrm{c}}} |\alpha_{0k}| \ge c_\alpha$, and $\min_{k\in\mathcal{A}_{\mathrm{v}}} \|\beta_{0k}\|_{L^2(\Omega)} \ge c_\beta$.

Assumption 10 (Covariates). There exists a constant c_X such that $|X_k| < c_X$ with probability one, for $k = 1, \ldots, p$.

Assumption 11 (Errors). The $\varepsilon_i(\cdot)$ is independent square-integrable random process, and $\boldsymbol{\varepsilon}_i$ is sub-Gaussian with variance proxy σ^2, that is, $\mathrm{E}\{\exp(\boldsymbol{t}^\top \boldsymbol{\varepsilon}_i)\} \le \exp(\sigma^2/2)$, for any $\|\boldsymbol{t}\|^2 = 1$ and some $\sigma > 0$. Each component of $\{\eta(\boldsymbol{s}) : \boldsymbol{s} \in \Omega\}$, $\{\eta(\boldsymbol{s})\eta(\boldsymbol{s}')^\top : (\boldsymbol{s}, \boldsymbol{s}') \in \Omega^2\}$ and $\{\boldsymbol{X}\eta(\boldsymbol{s}) : \boldsymbol{s} \in \Omega\}$ are Donsker classes.

Assumption 12 (Initial estimators). The initial estimators satisfy that $r_{n\alpha} \max_{k \notin \mathcal{A}_c} |\widetilde{\alpha}_k| = O_P(1)$, $r_{n\gamma} \max_{k \notin \mathcal{A}_v} \|\widetilde{\boldsymbol{\gamma}}_k\| = O_P(1)$, $r_{n\alpha}$, $r_{n\gamma} \to \infty$; and there exist positive constants b_α and b_γ such that $\Pr(\min_{k \in \mathcal{A}_c} |\widetilde{\alpha}_k| \ge c_\alpha b_\alpha) \to 1$ and $\Pr(\min_{k \in \mathcal{A}_v} \|\widetilde{\boldsymbol{\gamma}}_k\| \ge c_\gamma b_\gamma) \to 1$.

Assumption 13 (Parameters for consistency). Suppose that

$$\frac{\sqrt{nN_s^2 \log(p)}}{\rho_{n1} r_{n\alpha}} + \frac{\sqrt{nN_s^2 \log(pJ_n)}}{\rho_{n2} r_{n\gamma}} + \frac{nN_s|\triangle|^{d+1}}{\rho_{n1} r_{n\alpha} \wedge \rho_{n2} r_{n\gamma}} = o(1).$$

$$\frac{\rho_{n1}^2 + \rho_{n2}^2}{nN_s^2} = o(1), \quad \frac{n}{J_n^{(d+1)} \log(pJ_n)} = o(1).$$

Assumptions 9–11 are regularity conditions that are commonly used in the high-dimension literature. To obtain the selection consistency of the proposed method, we need an order requirement for a general initial estimator; see Assumption 12. Theorem S.1 in the supplementary materials demonstrates that the group LASSO estimator (Yuan and Lin 2006) obtained from (S.20) satisfies Assumption 12 under some weak conditions; specifically, if the tuning parameters satisfy $\widetilde{\rho}_{n1}^2 + \widetilde{\rho}_{n2}^2 \asymp nN_s^2 \log(p)$, then the consistent rates for the group LASSO estimator in Assumption 12 have order $r_{n\alpha} \asymp r_{n\gamma} = O\{\sqrt{n/\log(p)}\}$. Thus, we recommend using the group LASSO estimator as our initial estimator. Consequently, Assumption 13 indicates that $p = \exp\{o(n^{1/2})\}$.

The following theorems establish the asymptotic properties of the adaptive group LASSO estimators. Theorem 4 shows that the proposed procedure is consistent in both variable selection and the separation of varying and constant coefficients. Theorem 5 establishes the convergence rates of the estimators. Theorem 6 presents the asymptotic normality of the constant-coefficient estimator.

Theorem 4. Suppose that Assumptions 1–6, 9–13, and model (14) hold. Then for $\widehat{\mathcal{A}}_c$, $\widehat{\mathcal{A}}_v$ in (17), as $N_s \to \infty$ and $n \to \infty$, $\Pr(\widehat{\mathcal{A}}_c = \mathcal{A}_c) \to 1$, $\Pr(\widehat{\mathcal{A}}_v = \mathcal{A}_v) \to 1$.

Theorem 5. Under model (14) and Assumptions 1–6, 9–13, we have

$$\begin{aligned}
&\sum_{k \in \mathcal{A}_c} (\widehat{\alpha}_k - \alpha_{0k})^2 \\
&= O_P\left\{\frac{1}{n} + \frac{1}{nN_s|\triangle|^2} + |\triangle|^{2(d+1)} + \frac{\rho_{n1}^2 + \rho_{n2}^2}{n^2 N_s^2}\right\}, \\
&\sum_{k \in \mathcal{A}_v} \|\widehat{\beta}_k - \beta_{0k}\|_{L_2(\Omega)}^2 \\
&= O_P\left\{\frac{1}{n} + \frac{1}{nN_s|\triangle|^2} + |\triangle|^{2(d+1)} + \frac{\rho_{n1}^2 + \rho_{n2}^2}{n^2 N_s^2}\right\}.
\end{aligned}$$

Theorem 6. Under model (14) and Assumptions 1–8, 9–13, as $N_s \to \infty$ and $n \to \infty$, the estimator $\widehat{\boldsymbol{\alpha}}_{\mathcal{A}_c}$ satisfies that $n^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{\mathcal{A}_c} - \boldsymbol{\alpha}_{0,\mathcal{A}_c}\right) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Xi}^{-1}\boldsymbol{\Xi}_e\boldsymbol{\Xi}^{-1})$, where $\boldsymbol{\Xi}$ and $\boldsymbol{\Xi}_e$ are given in (11).

4. Numerical Studies

4.1. Triangulation, d, r, and Tuning Parameter Selection

To apply the bivariate spline smoothing, we need to choose the triangulation, a notoriously difficult task for nonparametric smoothing methods. An optimal triangulation is a partition of the domain which is best according to some criterion that measures the shape, size or number of triangles. For example, a "good" triangulation usually refers to those with well-shaped triangles, no small angles or/and no obtuse angles. Other criteria include the density control (adaptivity) and optimal size (number of triangles), etc. For a fixed number of triangles, Lai and Schumaker (2007) recommended selecting the triangulation according to max-min criterion which maximizes the minimum angle of all the angles of the triangles in the triangulation. In our theoretical studies, we find that the number of bivariate spline basis functions J_n needs to satisfy Assumption 6 in Section 2.2 and Assumption 13 in Section 3.2. Therefore, in practice, if the boundary of the spatial domain is not too complicated, we suggest taking the number of triangles N as the following: $N = \min\{\lfloor c_1 n^{1/(2d+2)} N_s^{1/2} \rfloor, N_s/10\} + c_2$, for tuning parameter c_1 (typically, $c_1 \in [0.3, 1]$ and $c_2 = 10$ work very well in practice). Once N is chosen, one can build the triangulated meshes using typical triangulation construction methods such as Delaunay triangulation.

Next, we discuss how to choose the smoothness r and degree d for the bivariate splines. For estimation, the proposed spline method with a higher degree is expected to result in a better estimation accuracy when the underlying coefficient function is smooth. According to Lai and Schumaker (2007), for a fixed $r \ge 1$, the bivariate spline achieves full approximation power asymptotically when $d \ge 3r + 2$. However, there are too many parameters to be estimated if d or r is too large, which may result in singularities and unnecessary computation burden. In practice, for some smooth functions without sharp edges as in the simulation studies, we suggest employing smoothness $r = 1$ with degree $d = 3$ or $d = 4$. On the other hand, if the underlying coefficient functions are believed to have sharp edges, we suggest using $r = 0$ and $d = 1$ or 2. For model structure identification, we find the proposed SMILE method is less sensitive to the choice of r and d compared to the estimation, so we recommend $r = 0$ or 1 with $d = 1$ or 2 to reduce the computational burden.

As for the selection of penalty parameters, we borrow the idea from Bayesian information criteria widely used in high-dimensional settings (Wang, Li, and Tsai 2007; Lee, Noh, and Park 2014). For any penalty parameter $\boldsymbol{\rho} = (\rho_1, \rho_2)^\top$, let $\mathrm{MSE}_{\boldsymbol{\rho}} = (nN_s)^{-1} \sum_{i=1}^n \sum_{j=1}^{N_s} (Y_{ij} - \widehat{\mu}_{\boldsymbol{\rho},ij})^2$ be the mean squared errors associated with $\boldsymbol{\rho}$, where $\widehat{\mu}_{\boldsymbol{\rho},ij} = \widehat{\mu}_{\boldsymbol{\rho},i}(\boldsymbol{s}_j)$ is defined in Remark 3. Let $\mathcal{I}_{\boldsymbol{\rho},c}$ and $\mathcal{I}_{\boldsymbol{\rho},v}$ be identified active constant and varying index set corresponding to tuning parameter $\boldsymbol{\rho}$, respectively. For the selection of the tuning penalty parameters, we modify the classical BIC rule in the following way to reflect

the functional data structure:

$$\mathrm{mBIC}^1_{\boldsymbol{\rho}} = \log\left(\mathrm{MSE}_{\boldsymbol{\rho}}\right) + \left\{\frac{|\mathcal{I}_{\boldsymbol{\rho},\mathrm{c}}|\log(n)}{n} + \frac{|\mathcal{I}_{\boldsymbol{\rho},\mathrm{v}}|J_n\log(nN_s)}{nN_s}\right\},$$

$$\mathrm{mBIC}^2_{\boldsymbol{\rho}} = \log\left(\mathrm{MSE}_{\boldsymbol{\rho}}\right) + \left\{\frac{|\mathcal{I}_{\boldsymbol{\rho},\mathrm{c}}|\log(p)\times\log(n)}{2n} + \frac{|\mathcal{I}_{\boldsymbol{\rho},\mathrm{v}}|J_n\log(pJ_n)\times\log(nN_s)}{2nN_s}\right\}.$$

Theorem S.2 in the supplementary materials shows the consistency of the $\mathrm{mBIC}^1_{\boldsymbol{\rho}}$ and $\mathrm{mBIC}^2_{\boldsymbol{\rho}}$ under some additional regularity assumptions. In our Monte Carlo simulation studies, we find that these two criteria give very similar results.

4.2. Simulation Study 1

In this simulation study, we evaluate the performance of the "oracle" PLSVCM estimator, that is, the estimator proposed in Section 2 when $\mathcal{A}_\mathrm{c}$ and $\mathcal{A}_\mathrm{v}$ in model (2) are known. We generate data at a grid of $N_s = n_1 \times n_2$ voxels for n subjects. At each $\boldsymbol{s} = (s_1, s_2)^\top$ in Ω, $Y_i(\boldsymbol{s})$ is simulated according to

$$Y_i(\boldsymbol{s}_j) = \sum_{k\in\mathcal{A}_\mathrm{c}} X_{ik}\alpha_k + \sum_{k\in\mathcal{A}_\mathrm{v}} X_{ik}\beta_k(\boldsymbol{s}_j) + \eta_i(\boldsymbol{s}_j) + \varepsilon_i(\boldsymbol{s}_j),$$

for $i = 1,\ldots,n$ and $j = 1,\ldots,N_s$. We take $n = 50$, 100, and 200, $n_1 = n_2 = 40$ and 50, $\mathcal{A}_\mathrm{c} = \{1,2\}$, and $\mathcal{A}_\mathrm{v} = \{3,4\}$. We generate $X_{i1},\ldots,X_{i4}$ independently from $\mathrm{Unif}[-1,1]$. We consider two scenarios below to illustrate the efficiency of the proposed estimator for PLSVCM.

- Scenario 1: $\alpha_1 = \alpha_2 = 0$;
- Scenario 2: $\alpha_1 = 1$ and $\alpha_2 = -1$.

Note that the PLSVCM is reduced to SVCM in Scenario 1. In both scenarios, we generate nonzero varying functions in the quadratic and exponential forms, where $\beta_3(\boldsymbol{s}) = 20\{(s_1 - 0.5)^2 + (s_2 - 0.5)^2\}$, and $\beta_4(\boldsymbol{s}) = \exp[-15\{(s_1 - 0.5)^2 + (s_2 - 0.5)^2\}]$.

To simulate the within-image dependence, for $i = 1,\ldots,n$, we generate $\eta_i(\boldsymbol{s})$ from $\eta_i(\boldsymbol{s}) = \sum_{\kappa=1}^2 \sqrt{\lambda_\kappa}\xi_{i\kappa}\psi_\kappa(\boldsymbol{s})$, where for $\kappa = 1,2$, we generate $\xi_{ik} \overset{\mathrm{iid}}{\sim} N(0,1)$, $\psi_1(\boldsymbol{s}) = a_1\sin(\pi s_1)$, and $\psi_2(\boldsymbol{s}) = a_2\{\cos(\pi s_2) + a_0\}$. We take $a_0 = -0.039$, $a_1 = 1.588$, and $a_2 = 2.157$, so that $\int_\Omega \psi_1^2(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = \int_\Omega \psi_2^2(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = 1$ and $\int_\Omega \psi_1(\boldsymbol{s})\psi_2(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = 0$, and the eigenvalues are $\lambda_1 = 0.3^2$ and $\lambda_2 = 0.075^2$.

The errors $\{\varepsilon(\boldsymbol{s}) : \boldsymbol{s} \in \Omega\}$ are generated in Gaussian process with mean zero and standard deviation σ_ε. In addition, we adjust the values of σ_ε to ensure the signal-noise-ratio (SNR) is close to 3 and 5 in both scenarios, where the SNR is defined as the following:

$$\mathrm{SNR} = \frac{N_s^{-1}\sum_{j=1}^{N_s}\mathrm{var}\{\sum_{k\in\mathcal{A}_\mathrm{c}} X_{ik}\alpha_k + \sum_{k\in\mathcal{A}_\mathrm{v}} X_{ik}\beta_k(\boldsymbol{s}_j)\}}{N_s^{-1}\sum_{j=1}^{N_s}\mathrm{var}\{\eta_i(\boldsymbol{s}_j) + \varepsilon_i(\boldsymbol{s}_j)\}}.$$

To apply our method, we consider a triangulation containing 29 triangles with 28 vertices, and based on this triangulation, we generate the bivariate spline basis functions with degree $d = 3$ and smoothness parameter $r = 1$.

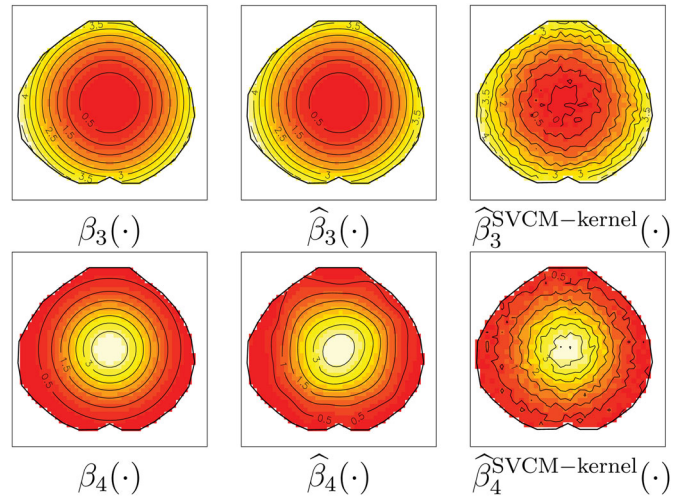


Figure 1. True and estimated varying coefficient functions. $\beta_3(\cdot)$ and $\beta_4(\cdot)$ are the true varying coefficient functions; $\widehat{\beta}_3(\cdot)$ and $\widehat{\beta}_4(\cdot)$ are the estimated varying coefficient functions by SVCM-BST; $\widehat{\beta}_3^{\mathrm{SVCM-kernel}}(\cdot)$ and $\widehat{\beta}_4^{\mathrm{SVCM-kernel}}(\cdot)$ are the estimates by SVCM-kernel.

Note that the underlying model for Scenario 1 is SVCM, and in this scenario we compare the proposed BST estimator (SVCM-BST) with the three stage kernel (SVCM-kernel) method proposed by Zhu, Fan, and Kong (2014). The true surfaces and the estimated functions from both methods are shown in Figure 1. Visually one can see that the estimated functions from the BST method successfully capture the spatial pattern of the true functions, while SVCM-kernel fails to capture the spatial structure and yield a smooth estimate. We evaluate the methods on the accuracy of coefficient estimation and precision. To be specific, we compute the mean squared errors (MSEs) to evaluate the accuracy of estimators, based on 100 Monte Carlo simulations. The MSEs for the coefficients functions $\beta_k(\cdot)$, $k = 3, 4$, are computed as the average of $N_s^{-1}\sum_{j=1}^{N_s}\{\widehat{\beta}_k(\boldsymbol{s}_j) - \beta_{0k}(\boldsymbol{s}_j)\}^2$.

Tables 1 and 2 present the MSEs in Scenarios 1 and 2, respectively. As shown in Tables 1 and 2, with the increase of sample size, the MSEs of all the estimators decrease significantly (all the MSE values at $n = 200$ is less than quarter of the values at $n = 50$). While with the increase of pixel size N_s, these values of MSEs almost keep constant. In addition, as shown in Table 1, the proposed SVCM-BST method outperforms the SVCM-kernel

Table 1. Mean squared error (MSE $\times 10^{-2}$) of the SVCM-BST and SVCM-kernel estimators of $\beta_k(\cdot)$, $k = 3, 4$, for Scenario 1 in Simulation Study 1.

			$\beta_3(\cdot)$		$\beta_4(\cdot)$	
N_s	n	SNR	SVCM-BST	SVCM-kernel	SVCM-BST	SVCM-kernel
40^2	50	3	1.864	3.352	2.623	3.550
		5	1.658	2.330	2.425	2.686
	100	3	1.028	2.105	0.961	1.595
		5	0.933	1.464	0.857	1.080
	200	3	0.489	1.277	0.627	1.117
		5	0.441	0.854	0.579	0.773
50^2	50	3	1.725	2.990	2.515	3.371
		5	1.598	2.075	2.381	2.550
	100	3	0.963	1.800	0.892	1.382
		5	0.902	1.275	0.826	0.958
	200	3	0.458	1.076	0.596	0.943
		5	0.427	0.724	0.565	0.673

Table 2. Mean squared error (MSE $\times 10^{-2}$) of the PLSVCM-BST estimator for Scenario 2 in Simulation Study 1.

		SNR = 3				SNR = 5			
N_s	n	α_1	α_2	$\beta_3(\cdot)$	$\beta_4(\cdot)$	α_1	α_2	$\beta_3(\cdot)$	$\beta_4(\cdot)$
40^2	50	1.605	1.559	2.077	2.839	1.608	1.553	1.843	2.616
	100	0.788	0.767	1.089	1.032	0.789	0.763	0.979	0.914
	200	0.387	0.297	0.511	0.649	0.386	0.297	0.456	0.594
50^2	50	1.588	1.534	1.907	2.707	1.593	1.533	1.762	2.558
	100	0.786	0.752	1.007	0.949	0.786	0.752	0.938	0.874
	200	0.383	0.294	0.473	0.611	0.382	0.295	0.438	0.576

in terms of MSE, regardless of the settings. The SVCM-kernel method requires the calculation of spatial similarity distance and adaptive weights repeatedly, and the estimate is constrained to uni-scale neighborhood; thus, it cannot efficiently estimate the coefficient functions. In contrast, our BST method provides a large amount of flexibility as it allows an easy implementation of piecewise polynomial representations of various degrees and smoothness over an arbitrary triangulation. As a result, it can handle irregularly shaped objects of different visual qualities.

As for the precision of the estimators, we assess the accuracy of the standard error (SE) formula of $\mathbf{V}_n^c$ given in (12), by calculating SE from three standards in Scenario 2, as shown in Table 3. $\mathrm{SE}_{\mathrm{mc}}$ can be viewed as the true values for SE, which is the standard deviation of the estimated parameters based on 100 Monte Carlo samples; $\mathrm{SE}_{\mathrm{mean}}$ and $\mathrm{SE}_{\mathrm{med}}$ are the mean and median of the estimated SE from 100 simulations. We use the $\mathrm{SE}_{\mathrm{IQR}}$ to assess the precision of the standard error formula; specifically, it is calculated by the interquartile range of the estimated SE from 100 Monte Carlo replications divided by 1.349.

As shown in Table 3, with the increase of sample size, all the three estimates of SE decreases significantly (all the SE values at $n = 200$ is less than half of the values at $n = 50$), indicating more precise estimates for constant coefficients. While with the increase of pixel size N_s, these values of SEs almost keep the same, indicating that the proposed method is stable in estimation. $\mathrm{SE}_{\mathrm{mean}}$ and $\mathrm{SE}_{\mathrm{med}}$ are very close to $\mathrm{SE}_{\mathrm{mc}}$; and with the increase of sample size, they get closer to the true SE. The values of the $\mathrm{SE}_{\mathrm{IQR}}$ are much smaller compared to those of the other three SEs, which implies the variance of the SE calculated by our formula is very small. Thus, it is safe to conclude that the proposed standard error formula provides a stable and accurate estimate.

4.3. Simulation Study 2

This simulation study is designed to evaluate finite-sample performance of the proposed sparse learning and model structure identification method. Specifically, we examine the performance of the double regularization method using an imaging genetic study, where we focus on learning the changes of genetic (G) effects under different environmental (E) conditions, that is, gene-environment (G×E) interaction. We simulate data at all voxels on a 20×20 phantom image for n subjects. At each $\boldsymbol{s} = (s_1, s_2)^\top$ in Ω, $Y_i(\boldsymbol{s})$ is simulated according to

$$\begin{aligned} Y_i(\boldsymbol{s}_j) &= \mathbf{G}_i^\top \boldsymbol{f}^{\mathrm{G}}(\boldsymbol{s}_j) + \mathbf{E}_i^\top \boldsymbol{f}^{\mathrm{E}}(\boldsymbol{s}_j) + (\mathbf{G} \times \mathbf{E})_i^\top \boldsymbol{f}^{\mathrm{G\times E}}(\boldsymbol{s}_j) \\ &\quad + \eta_i(\boldsymbol{s}_j) + \varepsilon_i(\boldsymbol{s}_j), \\ &\quad i = 1, \ldots, n,\ j = 1, \ldots, N_s, \end{aligned}$$

where $\mathbf{G}_i = (G_{i1}, \ldots, G_{ip_1})^\top$, $\boldsymbol{f}^{\mathrm{G}} = (f_1^{\mathrm{G}}, \ldots, f_{p_1}^{\mathrm{G}})^\top$ denotes the genetic factors potentially associated with response variation (e.g., coded SNP genotypes), $\mathbf{E}_i = (E_{i1}, \ldots, E_{ip_2})^\top$, $\boldsymbol{f}^{\mathrm{E}} = (f_1^{\mathrm{E}}, \ldots, f_{p_2}^{\mathrm{E}})^\top$ denotes the environmental factors (e.g., age, gender, and medical treatments) that may affect the imaging measures, and $(\mathbf{G} \times \mathbf{E})_i = ((\mathbf{G} \times \mathbf{E})_{i1}, \ldots, (\mathbf{G} \times \mathbf{E})_{i,p_1p_2})^\top$, $\boldsymbol{f}^{\mathrm{G\times E}} = (f_1^{\mathrm{G\times E}}, \ldots, f_{p_1p_2}^{\mathrm{G\times E}})^\top$ denotes the G×E interaction. Note that $p = p_1 + p_2 + p_1p_2$ in this study.

We take $n = 30$, 40, and 50, $p_1 = 5$, 100, 200, and 500, and $p_2 = 3$. As for the active index sets, we set $\mathcal{A}_{\mathrm{c}}^{\mathrm{G}} = \{1\}$, $\mathcal{A}_{\mathrm{v}}^{\mathrm{G}} = \{2\}$, $\mathcal{N}^{\mathrm{G}} = \{3, \ldots, p_1\}$, $\mathcal{A}_{\mathrm{c}}^{\mathrm{E}} = \{1\}$, $\mathcal{A}_{\mathrm{v}}^{\mathrm{E}} = \{2\}$, $\mathcal{N}^{\mathrm{E}} = \{3\}$, $\mathcal{A}_{\mathrm{c}}^{\mathrm{G\times E}} = \{1\}$, $\mathcal{A}_{\mathrm{v}}^{\mathrm{G\times E}} = \emptyset$, $\mathcal{N}^{\mathrm{G\times E}} = \{2, \ldots, p_1p_2\}$. The nonzero functions $f_1^{\mathrm{G}}(\boldsymbol{s})$, $f_1^{\mathrm{E}}(\boldsymbol{s})$, and $f_1^{\mathrm{G\times E}}(\boldsymbol{s})$ are linear, and $f_2^{\mathrm{G}}(\boldsymbol{s})$ and $f_2^{\mathrm{E}}(\boldsymbol{s})$ are varying functions; that is, $f_1^{\mathrm{G}}(\boldsymbol{s}) = \alpha_1^{\mathrm{G}}$, $f_2^{\mathrm{G}}(\boldsymbol{s}) = \beta_2^{\mathrm{G}}(\boldsymbol{s})$, $f_1^{\mathrm{E}}(\boldsymbol{s}) = \alpha_1^{\mathrm{E}}$, $f_2^{\mathrm{E}}(\boldsymbol{s}) = \beta_2^{\mathrm{E}}(\boldsymbol{s})$, and $f_1^{\mathrm{G\times E}}(\boldsymbol{s}) = \alpha_1^{\mathrm{G\times E}}$. We take $\alpha_1^{\mathrm{G}} = 2$, $\alpha_1^{\mathrm{E}} = 1$, and $\alpha_1^{\mathrm{G\times E}} = 1.5$. The varying functions and the within-image errors $\{\eta(\boldsymbol{s}) : \boldsymbol{s} \in \Omega\}$ are randomly generated

Table 3. Standard errors of the constant coefficients estimator for Scenario 2 in Simulation Study 1.

			α_1				α_2			
N_s	n	SNR	$\mathrm{SE}_{\mathrm{mc}}$	$\mathrm{SE}_{\mathrm{mean}}$	$\mathrm{SE}_{\mathrm{med}}$	$\mathrm{SE}_{\mathrm{IQR}}$	$\mathrm{SE}_{\mathrm{mc}}$	$\mathrm{SE}_{\mathrm{mean}}$	$\mathrm{SE}_{\mathrm{med}}$	$\mathrm{SE}_{\mathrm{IQR}}$
40^2	50	3	0.1253	0.1214	0.1211	0.0118	0.1240	0.1218	0.1215	0.0117
		5	0.1254	0.1212	0.1208	0.0115	0.1238	0.1216	0.1215	0.0116
	100	3	0.0885	0.0856	0.0858	0.0053	0.0876	0.0856	0.0855	0.0060
		5	0.0886	0.0856	0.0857	0.0055	0.0873	0.0855	0.0854	0.0062
	200	3	0.0610	0.0607	0.0604	0.0029	0.0540	0.0607	0.0604	0.0030
		5	0.0609	0.0606	0.0603	0.0029	0.0541	0.0606	0.0604	0.0029
50^2	50	3	0.1245	0.1208	0.1208	0.0113	0.1230	0.1212	0.1215	0.0111
		5	0.1247	0.1207	0.1206	0.0114	0.1230	0.1211	0.1216	0.0111
	100	3	0.0884	0.0854	0.0855	0.0058	0.0867	0.0853	0.0854	0.0064
		5	0.0884	0.0853	0.0854	0.0057	0.0867	0.0852	0.0853	0.0063
	200	3	0.0608	0.0604	0.0601	0.0028	0.0538	0.0604	0.0602	0.0029
		5	0.0607	0.0604	0.0600	0.0028	0.0539	0.0604	0.0601	0.0028

$\mathrm{SE}_{\mathrm{mc}}$, the standard deviation of estimated parameters based on 100 Monte Carlo samples (can be viewed as the true values for SE); $\mathrm{SE}_{\mathrm{mean}}$, mean of the estimated SE from 100 simulations; $\mathrm{SE}_{\mathrm{med}}$, median of the estimated SE from 100 simulations; $\mathrm{SE}_{\mathrm{IQR}}$, interquartile range of the estimated SE from 100 Monte Carlo replications divided by 1.349.

from Matérn processes with covariance

$$C(\|\boldsymbol{s}-\boldsymbol{s}'\|;\sigma_s^2,\theta,\nu)=\frac{\sigma_s^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\boldsymbol{s}-\boldsymbol{s}'\|}{\theta}\right)^{\nu}F_{\nu}\times\left(\frac{\sqrt{2\nu}\|\boldsymbol{s}-\boldsymbol{s}'\|}{\theta}\right),$$

where Γ is the gamma function, F_ν is the modified Neumann function, σ_s^2 is the sill, and θ is the range parameter. The Matérn process is very flexible family of stationary processes which produce more realistic structures for biological applications, as compared to Brownian motion or simple low dimensional structures. In our simulation, the nonzero varying functions are generated with parameters ($\sigma_s=1,\theta=1/4,\nu=5/2$), and $\eta(\boldsymbol{s})$ is generated in the same way but with parameters ($\sigma_s=1,\theta=1/4,\nu=3/2$), which results in the within-image errors less smooth than the parameter functions. The errors $\{\varepsilon(\boldsymbol{s}):\boldsymbol{s}\in\Omega\}$ are generated in Gaussian process with mean zero and standard deviation 1.5 and 2.2.

The covariates are simulated as follows. First, we generate $\widetilde{G}_{ik}$, $k=1,\ldots,p_1$, U_i, and $\widetilde{U}_i$ independently from standard uniform distribution, $i=1,\ldots,n$. Then we set $G_{ik}^0=(\widetilde{G}_{ik}+tU_i)/(1+t)$ for $k\in\mathcal{A}_{\mathrm{c}}^{\mathrm{G}}\cup\mathcal{A}_{\mathrm{v}}^{\mathrm{G}}$, and $G_{ik}^0=(\widetilde{G}_{ik}+t\widetilde{U}_i)/(1+t)$ for $k\notin\mathcal{A}_{\mathrm{c}}^{\mathrm{G}}\cup\mathcal{A}_{\mathrm{v}}^{\mathrm{G}}$, where the parameter t controls the correlation. To be specific, $\mathrm{corr}(G_{ik}^0,G_{ik'}^0)=t^2/(1+t^2)$ for $k,k'\in\mathcal{A}_{\mathrm{c}}^{\mathrm{G}}\cup\mathcal{A}_{\mathrm{v}}^{\mathrm{G}}$, $\mathrm{corr}(G_{ik}^0,G_{ik'}^0)=t^2/(1+t^2)$ for $k,k'\notin\mathcal{A}_{\mathrm{c}}^{\mathrm{G}}\cup\mathcal{A}_{\mathrm{v}}^{\mathrm{G}}$, and the active and inactive covariates are independent. Here we consider $t=1$. This generating method of G_{ik}^0 is the same as Example 1 of Huang, Horowitz, and Wei (2010). To mimic the genetic factors in real data, we set $G_{ik}=0$ if $G_{ik}^0\le 1/3$, $G_{ik}=1$ if $1/3<G_{ik}^0\le 2/3$, and $G_{ik}=2$ otherwise. Section S.9 of the supplementary materials provides another simulation study when all covariates in G part are correlated.

As for the $\mathbf{E}_i$, we generate E_{i1} independently from $N(0,1)$, E_{i2} independently from a Bernoulli distribution with success rate 0.5, and sample E_{i3} independently from $\{1,2,3\}$ with equal probability. In this example, the SNR is nearly 1 and 3, where the SNR is defined

$$\mathrm{SNR}=\frac{N_s^{-1}\sum_{j=1}^{N_s}\mathrm{var}\{\mathbf{G}_i^\top\boldsymbol{f}^{\mathrm{G}}(\boldsymbol{s}_j)+\mathbf{E}_i^\top\boldsymbol{f}^{\mathrm{E}}(\boldsymbol{s}_j)+(\mathbf{G}\times\mathbf{E})_i^\top\boldsymbol{f}^{\mathrm{G}\times\mathrm{E}}(\boldsymbol{s}_j)\}}{N_s^{-1}\sum_{j=1}^{N_s}\mathrm{var}\{\eta_i(\boldsymbol{s}_j)+\varepsilon_i(\boldsymbol{s}_j)\}}.$$

To approximate the varying coefficient functions, in the selection part, we use the normalized bivariate spline basis functions, with smoothness degree $d=2$, smoothness constraint $r=1$, and 29 triangles; and in the estimation part, to get better estimation results, we use the bases with $d=3$, $r=1$, and 29 triangles. For the selection of penalty parameters, we consider the modified BIC presented in Section 4.1. The simulation results are similar based on these criteria, so in the following, we choose ρ_1 by mBIC^1 and ρ_2 by mBIC^2 for illustration.

We adopt the following criteria to evaluate the methods on the accuracy of variable selection, model identification and prediction:

(B-i) The average percentage of nonzero constant coefficients correctly identified as nonzero constant ("TC");

(B-ii) The average percentage of nonzero varying coefficient functions correctly identified as nonzero varying functions ("TV");

(B-iii) The average percentage of true zero functions correctly identified as zero ("TN");

(B-iv) The average number of covariates with nonzero constant coefficients or nonzero varying functions incorrectly identified as having both zero linear and zero varying functions ("FN");

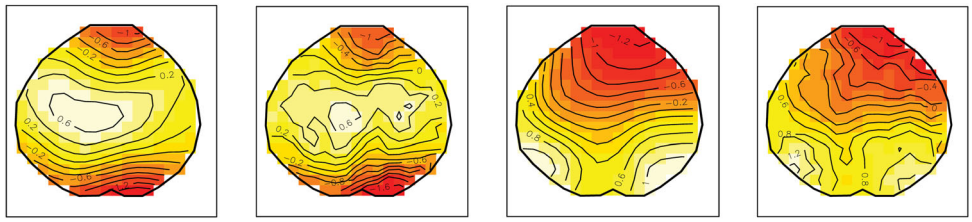
(C-i) Mean integrated squared errors ("MISE"), defined as the average of $\sum_{i=1}^n\sum_{j=1}^{N_s}\{\widehat{\mu}_i(\boldsymbol{s}_j)-\mu_i(\boldsymbol{s}_j)\}^2/(nN_s)$ over 100 simulations, where $\mu_i(\cdot)$ is the conditional mean function for the ith subject, $i=1,\ldots,n$.

All these performance measures are computed based on 100 replicates. Note that Criteria (B-i)–(B-iii) measure the frequency of getting the correct model structure; Criterion (B-iv) measure the frequency of getting an incorrect model structure; Criterion (C-i) focuses on the estimation and prediction accuracy for the model components. To evaluate the estimation accuracy, we compare the proposed SVCM method (SMILE) with the oracle PLSVCM-BST estimator, which all active and inactive index sets are treated as known. In this simulation study, the oracle PLSVCM-BST estimator works as a benchmark for estimation comparison. In the setting of $p_1=5$, we also compare with the SVCM-BST estimator, which ignores the identification of active index set and potential constant structure, and estimates with all covariates in a varying form. Note that SVCM-BST estimator is not applicable for high-dimensional settings, such as $p_1=100$, 200, or 500.

The true surfaces and the estimated functions are shown in Figure 2. Visually one can see that the estimated functions successfully capture the spatial pattern of the true functions.

The model selection results are provided in Table 4. The SMILE method can effectively identify informative constant and varying components in genetic markers (G part) and environmental factors (E part) as well as correctly discover the constant and varying spatial structure for their interactive effects (G×E part), regardless of the dependence structure. The proportions of correctly selected nonzero constant and varying components, for G, E, and G×E parts, are very close to perfect conditions (nearly 100% for TC, TV, and TN) in most cases, except for the cases with really small sample size, weak signal and large dimension; and the proportions increase to 100% as n increases and/or the SNR increases. The numbers of incorrectly identified components approach to zero as n increases and/or the SNR increases. The selection performance of the proposed method is great even with moderately large sample size and SNR.

The estimation results are displayed in Table 5. Specifically, we present the MISEs for estimators of the conditional mean functions. The case with known active covariates ("ORACLE") is reported in each setting and serves as a gold standard. In the setting of $p_1=5$, the results of the SVCM-BST estimator is also presented. In this setting, the proposed SMILE performs much better than the SVCM-BST, as indicated by MISEs, and is much closer to ORACLE, regardless of the SNR level and sample size n. That means, compared to the SVCM-BST that puts all covariates into the model fitting, the proposed SMILE method is able to



$\beta_2^{\mathrm{G}}(\cdot)$ $\widehat{\beta}_2^{\mathrm{G}}(\cdot)$ $\beta_2^{\mathrm{E}}(\cdot)$ $\widehat{\beta}_2^{\mathrm{E}}(\cdot)$

Figure 2. True and estimated varying coefficient functions.

Table 4. Selection statistics of the proposed method in Simulation Study 2.

			G Part				E Part				G×E Part		
n	SNR	p_1	TC	TV	TN	FN	TC	TV	TN	FN	TC	TN	FN
30	1	5	99	100	100	0.01	93	100	100	0.07	96	100	0.01
		100	92	53	100	0.55	66	61	100	0.73	85	100	0.08
		200	85	28	100	0.87	36	36	100	1.28	63	99.992	0.15
		500	71	9	100	1.2	31	19	100	1.5	54	99.996	0.29
	3	5	100	96	100	0.04	98	100	100	0.01	99	100	0
		100	98	96	100	0.06	85	97	100	0.18	92	99.990	0.02
		200	92	88	99.995	0.2	59	92	100	0.49	78	99.983	0.08
		500	88	77	99.994	0.35	48	81	100	0.71	73	99.989	0.12
40	1	5	99	100	100	0.01	98	100	100	0.02	99	100	0.01
		100	100	97	100	0.03	83	99	100	0.18	95	99.997	0
		200	99	94	100	0.07	80	97	100	0.23	93	99.998	0.01
		500	93	75	100	0.32	65	78	100	0.57	83	99.998	0.07
	3	5	100	91	100	0.09	100	100	100	0	100	100	0
		100	100	100	100	0	94	100	100	0.06	97	99.993	0
		200	100	100	100	0	86	100	100	0.14	96	99.997	0
		500	96	98	99.998	0.06	78	98	100	0.24	91	99.996	0.04
50	1	5	100	100	100	0	100	100	100	0	100	100	0
		100	100	100	100	0	96	100	100	0.04	100	100	0
		200	99	100	100	0.01	93	100	100	0.07	99	99.998	0.01
		500	100	100	100	0	93	100	100	0.07	99	99.999	0
	3	5	100	100	100	0	100	100	100	0	100	100	0
		100	100	99	100	0.01	99	100	100	0.01	100	100	0
		200	100	100	100	0	98	100	100	0.02	100	99.998	0
		500	100	100	100	0	98	100	100	0.02	99	99.999	0

TC, the average percentage of nonzero constant coefficients correctly identified as nonzero constant; TV, the average percentage of nonzero varying coefficient functions correctly identified as nonzero varying functions; TN, the average percentage of true zero functions correctly identified as zero; FN, the average number of covariates with nonzero constant coefficients or nonzero varying functions incorrectly identified as having both zero linear and zero varying functions.

Table 5. Mean integrated squared error (MISE) of proposed SMILE estimator, the unpenalized SVCM-BST estimator and the ORACLE for the mean function in Simulation Study 2.

		SNR = 1			SNR = 3		
n	p_1	SMILE	SVCM-BST	ORACLE	SMILE	SVCM-BST	ORACLE
30	5	0.314	2.064	0.207	0.164	1.318	0.141
	100	1.190	–	0.208	0.442	–	0.142
	200	2.029	–	0.207	0.931	–	0.141
	500	2.913	–	0.220	1.330	–	0.152
40	5	0.197	1.550	0.154	0.122	0.988	0.104
	100	0.364	–	0.158	0.194	–	0.108
	200	0.416	–	0.156	0.239	–	0.106
	500	1.012	–	0.155	0.484	–	0.106
50	5	0.125	1.245	0.125	0.086	0.798	0.086
	100	0.155	–	0.124	0.094	–	0.084
	200	0.223	–	0.127	0.101	–	0.086
	500	0.183	–	0.127	0.105	–	0.088

The SVCM-BST method with a fixed dimension is not applicable for $p_1 = 100$, 200, or 500, and for those cases "–" is used to indicate that results are not available.

correctly identify the active set and the model structure and lead to more accurate estimation for the conditional mean functions, even though it is subject to the bias of the variable selection. The results show the necessity of conducting variable selection and model identification, even in the case that $p < n$. When p_1 is much larger than the sample size n, the MISEs of the proposed method (SMILE) are slightly higher than those of the oracle estimators in the cases with moderate large sample size and/or SNR; while in the cases with small sample size, weak signal and large dimension, the estimates of the proposed SMILE has larger MISEs, restricted to the selection bias. Our estimates for the conditional mean functions are getting closer to the oracle estimators as sample size increases and the signal strengthens, which indicates the accuracy of the proposed method in estimation.

5. ADNI Data Analysis

We illustrate the application of our proposed method in the ultrahigh-dimensional setting by using the fludeoxyglucose positron emission tomography (PET) data of ADNI, which is publicly available and accessible through *http://adni.loni.usc.edu/*.

Single nucleotide polymorphisms (SNPs) and other polymorphisms in several genes, including apolipoprotein E

Table 6. Distribution of patients by sex and diagnosis status.

	Diagnosis status			
Sex	CN	MCI	AD	Total
Male	68	111	67	246
Female	39	44	45	128
Total	107	155	112	374

CN, cognitive normal; MCI, mild cognitive impairment; AD, Alzheimer's disease.

(APOE), have been demonstrated to be related to neuroimaging measures in brain disorders, such as mild cognitive impairment (MCI) and Alzheimer's disease (AD) (see, e.g., Ashford and Mortimer 2002). However, it is of great interest to identify other genes that play a role in the development and progression of MCI and AD. Here, we handle all the predictors (e.g., environmental/genetic factors and their interactions) jointly when investigating the association between imaging responses and scalar predictors. We develop an efficient method for G×E interaction identification to address the high-dimensionality of both the imaging and genomic data (Stein et al. 2010). Although interaction selection has drawn much attention in the literature (Hao and Zhang 2014; Kong et al. 2017; Li and Liu 2019), effectively relating hundreds of thousands of predictors to large-scale imaging data remains a challenging task.

This dataset is collected from 374 subjects. The data for each subject includes a spatially normalized PET image with 79×95 pixels, age, gender, cognitive impairment status, dummy variables of the number of copies of the epsilon 4 allele of APOE gene ($APOE_2$ as a dummy variable for subjects with two epsilon 4 alleles, and $APOE_1$ for those with one epsilon 4 allele), which is the strongest known genetic risk factor for AD, and other SNP genotypes at 620,901 loci. The original PET images are three-dimensional, and we focus on the 48th horizontal section of the brain from the bottom at the Corpus Callosum level. As stated in Marcus, Mena, and Subramaniam (2014), the two-dimensional image in this section cuts through the frontal and parietal lobes, which include many areas that may be altered in subjects with AD and/or dementia. The PET measurements ranged from -0.11 to 2.15, and the subjects' ages ranged from 56 to 89 years. Table 6 gives the distribution of patients by sex and diagnosis status.

We apply the proposed SVCM to capture important patient-level features that are associated with the variation of the PET images, identify the model structure (i.e., the features with varying or constant coefficients), and detect the association. We consider demographic variables including age and gender, dummy variables of AD and MCI, genetic information such as SNPs, and their interaction as possible features in our analysis.

To be specific, we first perform a nonparametric independence screening in VCMs (Fan, Ma, and Dai 2014) to SNP genotypes to reduce the computational burden, using piecewise constant bivariate splines over a triangulation with 1106 triangles and 596 vertices, as shown in the top-left corner of Figure 3. We select 40 SNPs that have the strongest associations with the image response conditional on age, gender, impairment status, and APOE genotype. Next, we apply our proposed SMILE method for (14) to model the relationship between the brain image response and 245 predictors, which consist of 40 selected SNP genotypes, along with age, gender, impairment status, APOE genotypes and their interactions. Note that for this setting, both of the SVCM-BST and SVCM-kernel fail in the regression, because of the memory limit. For selection, we use bivariate splines generated with degree $d = 2$ and smoothness parameter $r = 1$, and over a triangulation with 174 triangles and 28 vertices. Then we refit the model with those selected predictors using a finer triangulation with 1106 triangles and 596 vertices.

Table 8. Selected features with main or interaction effect for the ADNI data.

	× (Interaction)	(Main)	MCI	AD	Age	Sex	$APOE_1$	$APOE_2$
	(Main)		✓	✓	✓			✓
	rs1445493		✓					
	rs1541312	✓				✓		
	rs2131771	✓	✓				✓	
SNP	rs2955551					✓		
	rs7556318	✓	✓					
	rs8182037	✓	✓		✓	✓	✓	
	rs10935030	✓						
	rs10955341	✓				✓		

The "✓" indicates the existence of either main or interaction effect.

Tables 7 and 8 provide the information of the selected SNPs and covariates, respectively. Table 7 reports the chromosome information, physical position and gene information for selected SNPs with main or interaction effect. Table 8 lists the selected features with main or interaction effect for the ADNI data, where the checkmark indicates the existence of either main or interaction effect. For example, feature "age" has both main effect and interaction effect with SNP "rs8182037." The proposed SMILE method selects 10 main factors, including MCI, AD, age, $APOE_2$, and 6 SNPs; in addition, 11 G×E interaction factors are also selected.

All the 21 selected features are in spatially varying form, and the corresponding estimated coefficient functions are shown in Figure 3. The intercept image shows the PET image from a male normal individual, excluding effects from other features. The main effect of AD on PET image decreases 0.02–0.05 in the frontal lobe and parietal lobe. Compared to AD, the main effect of MCI has an increase of 0.1–0.15 in some spots in the frontal lobe near the longitudinal fissure. The main effect of age shows a decrease of 0.1 in every 20 years, in the frontal lobe and parietal lobe near the longitudinal fissure. $APOE_2$ has a different pattern with age and AD. There is an increase of 0.03–0.05 in the parietal lobe and frontal lobe near the longitudinal fissure. Most of the interaction terms with SNPs also show a decrease in the frontal lobe and/or parietal lobe, except for the interaction terms Sex by SNPs "rs8182037" and "rs10955341." These two terms show that female with high B allele frequency in "rs8182037" and "rs10955341" have an increase of 0.04–0.05 in the frontal lobe and parietal lobe.

6. Discussion

In this article, we consider the imaging-on-scalar regression with ultrahigh-dimensional scalar predictors. We have proposed a unified approach to perform estimation, variable selection and model structure identification simultaneously for bivariate functional responses with scalar covariates. FDA

Table 7. The chromosome information, physical position and gene information for selected SNPs with main or interaction effect.

SNP	CHR	BP	Gene	SNP	CHR	BP	Gene
rs1445493	8	9018773		rs7556318	1	87422453	
rs1541312	22	21000081	BMS1P20	rs8182037	15	73921656	UBE2Q2
rs2131771	11	12210124	MICAL2	rs10935030	3	133929794	NPHP3-AS1
rs2955551	8	8160844		rs10955341	8	105229526	RIMS2

CHR, chromosome; BP, physical position.

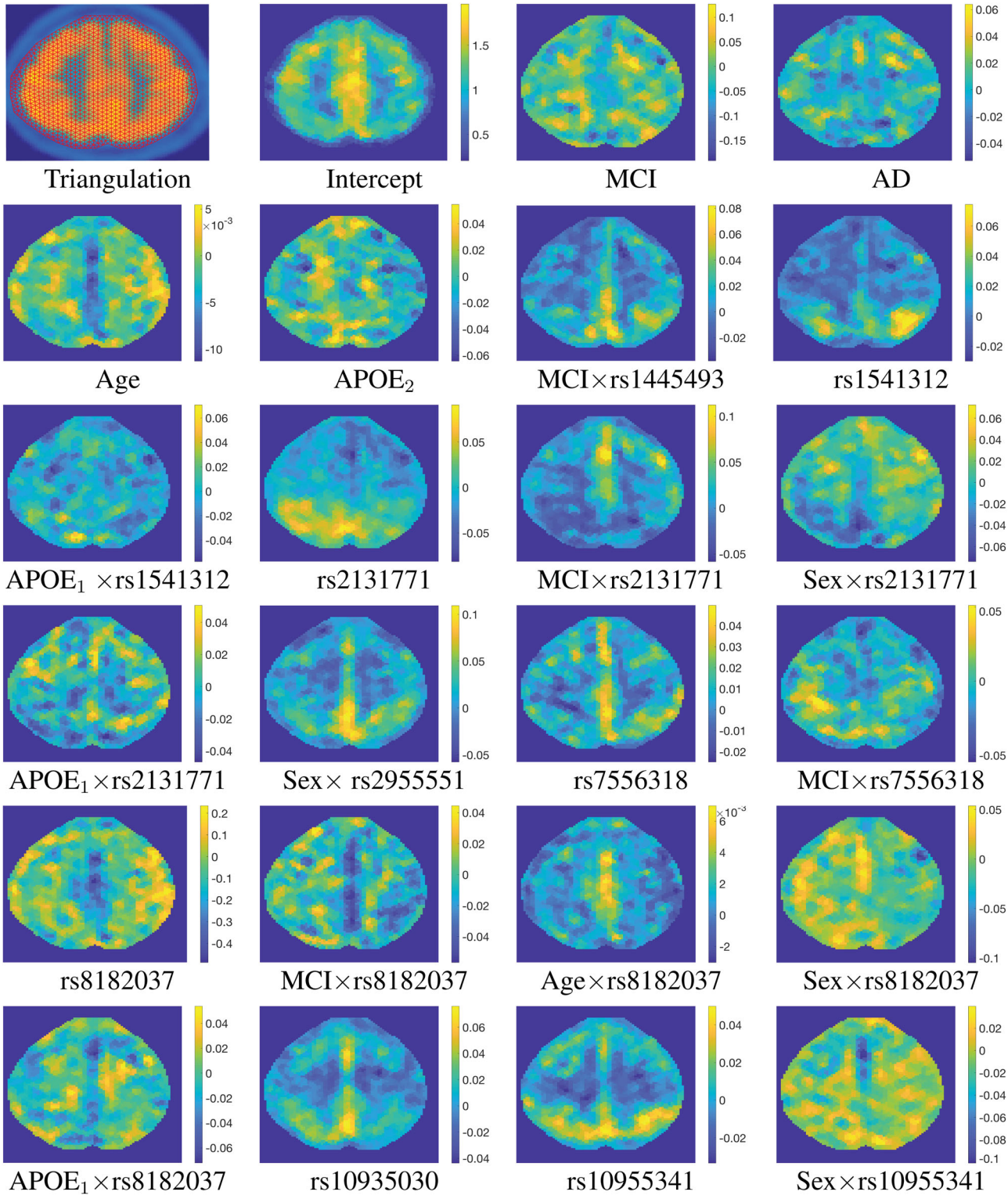


Figure 3. The estimated coefficient functions of features for the ADNI data.

methods have been widely studied in the literature, however, most focus on univariate functional responses, or bivariate functional responses (on rectangular domain) with fixed number of predictors. Little has been done on the development of advanced statistical methods to analyze complex functional data in high-dimensional setting. Thus, our work is novel, and it also has merits in the following aspects.

From a methodology point of view, our proposed procedure greatly enhances the capability of FDA in modern applications such as imaging genetics. We have successfully addressed the following challenges when analyzing next-generation functional data with complex data objects: (i) spatial smoothing over complex domains, (ii) high-dimensional functional responses, (iii) ultrahigh-dimensional covariates (such as high-throughput genotyping). Our article also enriches the methodological studies on cutting edge smoothing and estimation techniques to solve the problem of "leakage" in difficult regions.

In terms of theoretical innovations, we have shown that the proposed method has several desirable statistical properties. When the active constant varying index sets are known, we have established the consistency and asymptotic normality of the estimators for the coefficient parameters. We have also provided an asymptotic formula for estimating the standard errors of the coefficient estimates. With the availability of the asymptotic normality, one can easily devise a Wald-type of testing procedure for the coefficient parameters. In addition, we have derived model selection consistency for the proposed method and shown that it possesses the oracle property when the dimension of covariates exceeds the sample size. In addition, we have shown the estimators of the constant and varying coefficient are asymptotically normal under some regularity conditions.

In this article, we have conducted the GEWAS in neuroimaging genetics, where changes of genetic (G) effects under different environmental (E) conditions, that is, the G×E interaction, has been investigated. Thus, from an application point of view, our work has a direct impact on those biomedical imaging studies. In addition, the proposed SMILE method inherits the computational efficiency of the global smoothing methods, so that we can deal with large datasets with thousands of predictors, high-dimensional imaging data and a potentially large number of subjects within minutes.

A few more issues still merit further research. For instance, our method can be immediately extended to accommodate more complex data structures, such as three-dimensional images, by considering a functional linear regression model with trivariate functional responses, to fulfill the usefulness of the whole three-dimensional spatial information. Due to technical reasons, imaging data often have different modalities, for example, functional magnetic resonance imaging, diffusion tensor imaging, PET, computed tomography, ultrasonic imaging, and X-ray. It is interesting to examine the effect of the degree parameter d and smoothness parameter r of the BST smoothing on the performance of the proposed method for different types of imaging data.

Supplementary Materials

In the supplemental materials, we provide the technical proofs for the main theorems and additional simulation studies.

Acknowledgments

Data used in preparation of this article were obtained from the 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

This research is supported by National Science Foundation awards DMS-1542332 and DMS-1916204 (Li Wang) and the IR/D program from the National Science Foundation (Huixia Judy Wang). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Ahmad, I., Leelahanon, S., and Li, Q. (2005), "Efficient Estimation of a Semiparametric Partially Linear Varying Coefficient Model," *The Annals of Statistics*, 33, 258–283. [2]

Ashford, J. W., and Mortimer, J. A. (2002), "Non-Familial Alzheimer's Disease Is Mainly Due to Genetic Factors," *Journal of Alzheimer's Disease*, 4, 169–177. [12]

Chen, Y., Bai, Y., and Fung, W. (2017), "Structural Identification and Variable Selection in High-Dimensional Varying-Coefficient Models," *Journal of Nonparametric Statistics*, 29, 258–279. [6]

Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1057. [2]

Fan, J., Ma, Y., and Dai, W. (2014), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models," *Journal of the American Statistical Association*, 109, 1270–1284. [12]

Fotheringham, A., Brunsdon, C., and Charlton, M. (2002), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Hoboken, NJ: Wiley. [2]

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003), "Spatial Modeling With Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396. [2]

Gu, L., Wang, L., Härdle, W. K., and Yang, L. (2014), "A Simultaneous Confidence Corridor for Varying Coefficient Regression With Sparse Functional Data," *Test*, 23, 806–843. [2]

Hao, N., and Zhang, H. H. (2014), "Interaction Screening for Ultrahigh-Dimensional Data," *Journal of the American Statistical Association*, 109, 1285–1301. [12]

Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society*, Series B, 55, 757–796. [2]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *Annals of Statistics*, 38, 2282–2313. [10]

Huang, J., Wu, C. O., and Zhou, L. (2004), "Polynomial Spline Estimation and Inference for Varying Coefficient Models With Longitudinal Data," *Statistica Sinica*, 14, 763–788. [2,5]

Jiang, Q., Wang, H., Xia, Y., and Jiang, G. (2013), "On a Principal Varying Coefficient Model," *Journal of the American Statistical Association*, 108, 228–236. [2]

Kong, Y., Li, D., Fan, Y., and Lv, J. (2017), "Interaction Pursuit in High-Dimensional Multi-Response Regression via Distance Correlation," *The Annals of Statistics*, 45, 897–922. [12]

Lai, M. J., and Schumaker, L. L. (2007), *Spline Functions on Triangulations*, Cambridge: Cambridge University Press. [2,3,7]

Lee, E. R., Noh, H., and Park, B. U. (2014), "Model Selection via Bayesian Information Criterion for Quantile Regression Models," *Journal of the American Statistical Association*, 109, 216–229. [7]

Li, D., Ke, Y., and Zhang, W. (2015), "Model Selection and Structure Specification in Ultra-High Dimensional Generalised Semi-Varying Coefficient Models," *The Annals of Statistics*, 43, 2676–2705. [6]

Li, Y., and Liu, J. S. (2019), "Robust Variable and Interaction Selection for Logistic Regression and General Index Models," *Journal of the American Statistical Association*, 114, 271–286. [12]

Lian, H., Lai, P., and Liang, H. (2013), "Partially Linear Structure Selection in Cox Models With Varying Coefficients," *Biometrics*, 69, 348–357. [6]

Lian, H., Meng, J., and Zhao, K. (2015), "Spline Estimator for Simultaneous Variable Selection and Constant Coefficient Identification in High-Dimensional Generalized Varying-Coefficient Models," *Journal of Multivariate Analysis*, 141, 81–103. [6]

Liu, J., Li, R., and Wu, R. (2014), "Feature Selection for Varying Coefficient Models With Ultrahigh-Dimensional Covariates," *Journal of the American Statistical Association*, 109, 266–274. [2]

Lu, Z., Steinskog, D. J., Tjøstheim, D., and Yao, Q. (2009), "Adaptively Varying-Coefficient Spatiotemporal Models," *Journal of the Royal Statistical Society*, Series B, 71, 859–880. [2]

Marcus, C., Mena, E., and Subramaniam, R. M. (2014), "Brain PET in the Diagnosis of Alzheimer's Disease," *Clinical Nuclear Medicine*, 39, e413–e426. [12]

Morris, J. S. (2015), "Functional Regression," *Annual Reviews of Statistics and Its Application*, 2, 321–359. [1]

Morris, J. S., and Carroll, R. J. (2006), "Wavelet-Based Functional Mixed Models," *Journal of the Royal Statistical Society*, Series B, 68, 179–199. [2]

Mu, J., Wang, G., and Wang, L. (2018), "Estimation and Inference in Spatially Varying Coefficient Models," *Environmetrics*, 29, e2485. [2]

Noh, H., Chung, K., and Van Keilegom, I. (2012), "Variable Selection of Varying Coefficient Models in Quantile Regression," *Electronic Journal of Statistics*, 6, 1220–1238. [6]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, New York: Springer. [1]

Ramsay, T. (2002), "Spline Smoothing Over Difficult Regions," *Journal of the Royal Statistical Society*, Series B, 64, 307–319. [2]

Reiss, P. T., and Ogden, R. T. (2010), "Functional Generalized Linear Models With Images as Predictors," *Biometrics*, 66, 61–69. [2]

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., and Huentelman, M. J. (2010), "Voxelwise Genome-Wide Association Study (vGWAS)," *Neuroimage*, 53, 1160–1174. [12]

Tang, Y., Wang, H. J., Zhu, Z., and Song, X. (2012), "A Unified Variable Selection Approach for Varying Coefficient Models," *Statistica Sinica*, 22, 601–628. [2]

Wang, D., and Kulasekera, K. (2012), "Parametric Component Detection and Variable Selection in Varying-Coefficient Partially Linear Models," *Journal of Multivariate Analysis*, 112, 117–129. [6]

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [7]

Wang, H., and Ranalli, M. G. (2007), "Low-Rank Smoothing Splines on Complicated Domains," *Biometrics*, 63, 209–217. [2]

Wang, H. J., Zhu, Z., and Zhou, J. (2009), "Quantile Regression in Partially Linear Varying Coefficient Models," *The Annals of Statistics*, 37, 3841–3866. [2]

Wang, J.-L., Chiou, J.-M., and Muller, H.-G. (2016), "Functional Data Analysis," *Annal Review of Statistics and Its Applications*, 3, 257–295. [1]

Wang, L., Wang, G., Lai, M. J., and Gao, L. (2020), "Efficient Estimation of Partially Linear Models for Data on Complicated Domains by Bivariate Penalized Splines Over Triangulations," *Statistica Sinica*, 30, 347–369. [2]

Yu, S., Wang, G., Wang, L., Liu, C., and Yang, L. (2019), "Estimation and Inference for Generalized Geoadditive Models," *Journal of the American Statistical Association*. [2,3]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [7]

Zhu, H., Fan, J., and Kong, L. (2014), "Spatially Varying Coefficient Model for Neuroimaging Data With Jump Discontinuities," *Journal of the American Statistical Association*, 109, 1084–1098. [2,8]