stats

MDPI

Article

Confounder Adjustment in Shape-on-Scalar Regression Model: Corpus Callosum Shape Alterations in Alzheimer's Disease

Harshita Dogra, Shengxian Ding, Miyeon Yeon, Rongjie Liu and Chao Huang *

Department of Statistics, Florida State University, Tallahassee, FL 32306, USA; hdogra@fsu.edu (H.D.); nding2@fsu.edu (S.D.); myeon@fsu.edu (M.Y.); rliu3@fsu.edu (R.L.)
* Correspondence: chuang7@fsu.edu

Abstract: Large-scale imaging studies often face challenges stemming from heterogeneity arising from differences in geographic location, instrumental setups, image acquisition protocols, study design, and latent variables that remain undisclosed. While numerous regression models have been developed to elucidate the interplay between imaging responses and relevant covariates, limited attention has been devoted to cases where the imaging responses pertain to the domain of shape. This adds complexity to the problem of imaging heterogeneity, primarily due to the unique properties inherent to shape representations, including nonlinearity, high-dimensionality, and the intricacies of quotient space geometry. To tackle this intricate issue, we propose a novel approach: a shape-on-scalar regression model that incorporates confounder adjustment. In particular, we leverage the square root velocity function to extract elastic shape representations which are embedded within the linear Hilbert space of square integrable functions. Subsequently, we introduce a shape regression model aimed at characterizing the intricate relationship between elastic shapes and covariates of interest, all while effectively managing the challenges posed by imaging heterogeneity. We develop comprehensive procedures for estimating and making inferences about the unknown model parameters. Through real-data analysis, our method demonstrates its superiority in terms of estimation accuracy when compared to existing approaches.

Keywords: imaging heterogeneity; Alzheimer's disease; corpus callosum; square root velocity function; shape-on-scalar regression model

Citation: Dogra, H.; Ding, S.; Yeon, M.; Liu, R.; Huang, C. Confounder Adjustment in Shape-on-Scalar Regression Model: Corpus Callosum Shape Alterations in Alzheimer's Disease. *Stats* **2023**, *6*, 980–989. https://doi.org/10.3390/stats6040061

Academic Editor: Wei Zhu

Received: 10 September 2023
Revised: 26 September 2023
Accepted: 27 September 2023
Published: 28 September 2023

1. Introduction

Multi-site neuroimaging data integrative analysis is becoming of great interest so that establishing the relationship between imaging responses and covariates of interest from large-scale imaging studies, such as the Alzheimer's Disease Initiative (ADNI) study [1], can identify significant biomarkers for major neurological diseases, irrespective of any technical barriers. The need for imaging integration strategies arises as the differences in image acquisition protocols, experimental designs, and other unknown hidden factors could lead to invalid associations between biological variables of interest and false conclusions.

Some statistical integration techniques have been developed and applied to neuroimaging data harmonization. ComBat was first developed to remove unwanted variations caused by batches (with small samples) in gene expression data [2]. Recently, ComBat was applied to neuroimage data by shrinking the batch effect to the overall mean batch effect across voxels using the empirical Bayes framework. ComBat-GAM is an extension of ComBat where nonlinear trends of demographic features like age and sex are accommodated in the location-scale adjustment model by using a generalized additive model [3]. Surrogate variable analysis is another widely used harmonization technique to estimate unknown hidden factors in gene expression studies [4]. It was also applied to neuroimaging data that successfully identified brain disorder identification and predicted disease progression after removal of inter-site heterogeneity [5–9]. Although these strategies have been developed to

harmonize multi-site effect and increase power and reproducibility of statistical tests of various imaging data modalities, including diffusion tensor images [10], cortical thickness measurements [11,12], and functional magnetic resonance imaging (MRI) data [13,14], limited work has been conducted for imaging responses derived from the shape space. The shape is broadly defined to be a characteristic that is left after certain nuisance or shape-preserving transformations, such as rotations, translations, and scale, are removed [15–17], with the result that shape representation spaces are nonlinear, high-dimensional, and have quotient space geometry [18]. Therefore, the conventional normality assumptions often applied to imaging responses in existing image-on-scalar regression models cannot be directly extended to shape data. Furthermore, most existing harmonization approaches fail to provide inference tools to investigate the model uncertainty [2,3].

This paper aims to introduce a shape-on-scalar regression model that incorporates confounder adjustment. In particular, we leverage the square root velocity function to extract elastic shape representations, which are embedded within the linear Hilbert space of square integrable functions. Subsequently, we introduce a shape regression model aimed at characterizing the intricate relationship between elastic shapes and covariates of interest, all while effectively managing the challenges posed by imaging heterogeneity. We develop comprehensive procedures for estimating and making inferences about the unknown model parameters. To demonstrate our method, we consider the corpus callosum contour (CC) shape data from ADNI study and investigate its relationship with some covariates of interest. CC is the largest white matter tract containing millions of nerve fibers that connect the left cerebral hemisphere to the right hemisphere. It facilitates cognitive functions like attention, memory, learning, reasoning, planning, and problem solving. Studies on Alzheimer's disease (AD) consistently show that the CC undergoes notable changes in the early stages of the disease. These alterations in both anterior and posterior CC sections correlate with cognitive decline. Specific MRI measurements of the CC are associated with this cognitive deterioration. Overall, patterns of callosal deformation are emerging as significant biomarkers for diagnosing and tracking AD progression [19–23].

The paper is organized as follows. Section 2 introduces a shape-on-scalar regression model with confounder adjustment and outlines the estimation and inference procedures for the model parameters. In Section 3, we apply our method to the CC shape dataset in the ADNI study.

2. Methods

2.1. Preliminaries and Notations

We suppose that we observe both imaging data and covariates of interest, including clinical variables and demographic information, from n unrelated subjects. Instead of the whole brain image, we are interested in the contour of the planar CC. We let $\mathbf{L}_i$ be a $m \times 2$ matrix with m landmarks representing the contour of the ROI in $\mathbb{R}^2$. In addition, we let $\mathbf{X}$ be an $n \times p$ full column rank matrix of observed covariates, including the intercept. We let $\boldsymbol{e}^{\otimes 2} = \boldsymbol{e}\boldsymbol{e}^\top$ for any vector $\boldsymbol{e}$, $\mathbf{A} \otimes \mathbf{B}$ be the Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$, and $diag(\cdot)$ represent a diagonal matrix with all the elements on the diagonal.

2.2. Shape-on-Scalar Regression Model

Given the landmarks $\mathbf{L}_i$ from the contour of planar CC, we first derive the coordinate functions, $\boldsymbol{f}_i(t) \doteq (f_{i,1}(t), f_{i,2}(t))$ with $f_{i,1}(t)$ and $f_{i,2}(t)$ in the x-axis and the y-axis, respectively, in Kendall's shape space, where shapes are invariant to shape-preserving transformations, e.g., rotation, translation, and scaling. Specifically, we remove these nuisance transformations from the landmarks $\mathbf{L}_i$ via applying some preprocessing steps (details can be found in [24]). After that, as illustrated in [25], we derive the square root velocity function (SRVF) representations for the ith subject:

$$\boldsymbol{g}_i(t) : [0,1] \to \mathbb{R}^2,\ \boldsymbol{g}_i(t) = (g_{i,1}(t), g_{i,2}(t)),\ g_{i,j}(t) = \dot{f}_{i,j}(t)/\sqrt{|\dot{f}_{i,j}(t)|},\ j = 1,2. \quad (1)$$

According to the results in [25], if the functions in $\boldsymbol{f}_i(t)$ are absolutely continuous, then the corresponding SRVF representations are square integrable, i.e., $\boldsymbol{g}_i(t) \in \mathbb{L}^2([0,1], \mathbb{R}^2)$, $i = 1, \ldots, n$. Under this representation, we can further determine the optimal registration (or re-parameterization) group action for each SRVF representation, which corresponds to the following optimization problem:

$$\gamma_{i,*}(t) = \text{arginf}_{\gamma(t)\in\Gamma} \|\boldsymbol{\mu}(t) - (\boldsymbol{g}_i \circ \gamma(t))\sqrt{\dot{\gamma}(t)}\|, \; i = 1, 2, \cdots, n, \tag{2}$$

where $\boldsymbol{\mu}(t)$ is a template, such as the mean of $\{\boldsymbol{g}_i(t), i = 1, \cdots, n\}$, Γ includes all possible diffeomorphisms of $[0,1]$ that preserve the boundaries, i.e., $\Gamma = \{\gamma(t) : [0,1] \to [0,1] \mid \gamma(0) = 0, \gamma(1) = 1\}$, and the composition $\boldsymbol{g}_i \circ \gamma(t)$ is a re-parameterization of $\boldsymbol{g}_i(t)$. Then, we can obtain the aligned SRVF representations as follows:

$$\psi(g_{i,j}(t), \gamma_{i,*}(t)) = (g_{i,j} \circ \gamma_{i,*}(t))\sqrt{\dot{\gamma}_{i,*}(t)}, j = 1, 2, i = 1, \ldots, n. \tag{3}$$

To investigate the relationship between shape representations and some covariates of interest while handling the heterogeneity introduced by unobserved confounders, we consider the following shape-on-scalar regression model:

$$\boldsymbol{\Psi}(t) = \mathbf{X}\mathbf{B}(t) + \boldsymbol{\Lambda}(t) + \boldsymbol{\eta}(t) + \boldsymbol{\epsilon}(t), \tag{4}$$

where $\boldsymbol{\Psi}(t)$ is a $n \times 2$ matrix, including the shape representations, with the jth column $\boldsymbol{\Psi}_{\cdot j}(t) = (\psi(g_{1,j}(t), \gamma_{1,*}(t)), \ldots, \psi(g_{n,j}(t), \gamma_{n,*}(t)))^\top, j = 1, 2$. In practice, we assume that the shape responses are observed at n_v grid points, denoted as $t_1, \ldots, t_{n_v}$. For the functional coefficients, $\mathbf{B}(t) = (\boldsymbol{\beta}_1(t), \boldsymbol{\beta}_2(t))$ is a $p \times 2$ matrix that represents the primary effect associated with $\mathbf{X}$. For the unobserved terms, $\boldsymbol{\Lambda}(t) = (\boldsymbol{\Lambda}_{\cdot 1}(t), \boldsymbol{\Lambda}_{\cdot 2}(t))$ is a $n \times 2$ matrix that represents the functional hidden confounders related to $\mathbf{X}$, $\boldsymbol{\eta}(t) = (\boldsymbol{\eta}_{\cdot 1}(t), \boldsymbol{\eta}_{\cdot 2}(t))$ is a $n \times 2$ matrix, independent of $\mathbf{X}$, representing both subject-specific and location-specific spatial variability in the shape representations, and $\boldsymbol{\epsilon}(t) = (\boldsymbol{\epsilon}_{\cdot 1}(t), \boldsymbol{\epsilon}_{\cdot 2}(t))$ is a $n \times 2$ matrix including the measurement errors. It is assumed that each row in $\boldsymbol{\eta}(t)$ and $\boldsymbol{\epsilon}(t)$ is a mutually independent and identical copy of $\text{SP}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ and $\text{SP}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, respectively. Here, $\text{SP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a stochastic process with mean function $\boldsymbol{\mu}(t)$ and covariance function $\boldsymbol{\Sigma}(s, s')$. Moreover, $\boldsymbol{\Sigma}_\epsilon(t, t')$ is in the form of $\boldsymbol{\Omega}_\epsilon(t)\mathbf{1}(t = t')$, where $\boldsymbol{\Omega}_\epsilon(t)$ is a diagonal matrix and $\mathbf{1}(\cdot)$ is an indicator function.

2.3. Estimation Procedure

To estimate the coefficient functions $\mathbf{B}(t)$ and functional hidden confounders $\boldsymbol{\Lambda}(t)$ in (4), we follow three main steps: (i) orthogonal decomposition of functional hidden confounders; (ii) rank-q representation of functional hidden confounders; and (iii) bias correction of the estimated coefficient functions.

(i) Orthogonal decomposition of functional hidden confounders. We apply the orthogonal decomposition of $\boldsymbol{\Lambda}_{\cdot j}(t)$ onto the columns of $\mathbf{X}$ and reparametrize (4) as

$$\boldsymbol{\Psi}_{\cdot j}(t) = \mathbf{X}\boldsymbol{\beta}_j^*(t) + \boldsymbol{\Lambda}_{\cdot j}^*(t) + \boldsymbol{\eta}_{\cdot j}(t) + \boldsymbol{\epsilon}_{\cdot j}(t), \tag{5}$$

where $\boldsymbol{\beta}_j^*(t) = \boldsymbol{\beta}_j(t) + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Lambda}_{\cdot j}(t)$, $\boldsymbol{\Lambda}_{\cdot j}^*(t) = (\mathbf{I}_n - \mathbf{P}_X)\boldsymbol{\Lambda}_{\cdot j}(t)$, and $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Then, we can obtain the weighted least squares estimate of $\boldsymbol{\beta}_j^*(t)$ by using the local linear kernel smoothing method in [26,27]. We let $K_{h_\beta}(t) = |h_\beta|^{-1}K(h_\beta^{-1}t)$ and $\boldsymbol{z}_{h_\beta}(t' - t) = (1, (t' - t)/h_\beta)^\top$, where $K(\cdot)$ is the kernel function and h_β is the bandwidth matrix. Then, the WLS estimator of $\boldsymbol{\beta}_j^*(t)$ can be written as

$$\widehat{\boldsymbol{\beta}}_j^*(t) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \sum_{k=1}^{n_v} a_k(h_\beta, t)\boldsymbol{\Psi}_j(t_k), \tag{6}$$

where $a_k(h_\beta, t) = (1, \mathbf{0}_{1\times d})[\sum_{k=1}^{n_v} K_{h_\beta}(t_k - t)\boldsymbol{z}_{h_\beta}(t_k - t)^{\otimes 2}]^{-1} K_{h_\beta}(t_k - t)\boldsymbol{z}_{h_\beta}(t_k - t)$. The local linear estimator we obtain in (6) is a biased estimator. We correct this bias using the pre-asymptotic substitution method in [27]. We find the bias term by fitting a local cubic with a pilot bandwidth selected in (6). Moreover, we find the estimate of $\boldsymbol{\Lambda}_{\cdot j}(t)$. Finally, we can subtract the estimate $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Lambda}}_{\cdot j}(t)$ from $\widehat{\boldsymbol{\beta}}_j^*(t)$ to achieve the unbiased estimator $\widehat{\boldsymbol{\beta}}_j(t)$.

(ii) Rank-q representation of functional hidden confounders. In order to estimate $\boldsymbol{\Lambda}(t)$, we consider a rank-q representation of $\boldsymbol{\Lambda}^*(t)$, i.e., $\mathbf{Z}\mathbf{A}(t)$, where $\mathbf{Z}$ is a $n \times q$ loading matrix and $\mathbf{A}(t) = (\boldsymbol{\alpha}_1(t), \boldsymbol{\alpha}_2(t))$ is a $q \times 2$ matrix representing the corresponding effect. To estimate the loading matrix $\mathbf{Z}$, we first compute the residuals from the first step as $\boldsymbol{R}_{\cdot j}(t) = \boldsymbol{\Psi}_{\cdot j}(t) - \mathbf{X}\widetilde{\boldsymbol{\beta}}_j(t)$ where $\widetilde{\boldsymbol{\beta}}_j(t)$ is the bias-corrected version of $\widehat{\boldsymbol{\beta}}_j^*(t)$. Then, we write the extended residual matrix as a $n \times 2n_v$ matrix denoted by $\bar{\mathbf{R}} = (\boldsymbol{R}_{\cdot 1}(t_1), \ldots, \boldsymbol{R}_{\cdot 1}(t_{n_v}), \boldsymbol{R}_{\cdot 2}(t_1), \ldots, \boldsymbol{R}_{\cdot 2}(t_{n_v}))$. Given $\mathbf{X}$ and $\boldsymbol{\Lambda}(t)$, we can write the conditional expectation of $\bar{\mathbf{R}}$ as [28]

$$\mathbb{E}[\bar{\mathbf{R}} \mid \mathbf{X}, \boldsymbol{\Lambda}(t)] = \mathbf{Z}\bar{\mathbf{A}} + o_p(h_\beta^2), \tag{7}$$

where $\bar{\mathbf{A}} = (\boldsymbol{\alpha}_1(t_1), \ldots, \boldsymbol{\alpha}_1(t_{n_v}), \boldsymbol{\alpha}_2(t_1), \ldots, \boldsymbol{\alpha}_2(t_{n_v}))$. Therefore, the loading matrix $\mathbf{Z}$ can be estimated through the singular value decomposition (SVD) on the extended residual matrix $\bar{\mathbf{R}}$, i.e., $\bar{\mathbf{R}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top$. $\mathbf{U}$ and $\mathbf{V}$ represent the left and right singular vectors and $\boldsymbol{\Delta}$ represents the diagonal matrix containing ordered singular values of $\bar{\mathbf{R}}$. The first q columns of $\mathbf{U}$ represented as $\mathbf{U}_{1:q}$ are the estimators of linear combinations of columns of $\mathbf{Z}$ [28]. Then, there exists a $q \times q$ orthonormal matrix $\mathbf{Q}$ such that $\mathbf{U}_{1:q} = \mathbf{Z}\mathbf{Q} + o_p(1)$ and $\mathbf{Z}\boldsymbol{\alpha}_j(t) = \mathbf{U}_{1:q}\boldsymbol{\delta}_j(t)$, where $\boldsymbol{\delta}_j(t) = \mathbf{Q}^\top\boldsymbol{\alpha}_j(t)$, $j = 1, 2$.

(iii) Bias correction of the estimated coefficient functions. We treat the residual terms in Step (ii) as functional responses and derive the estimate of $\boldsymbol{\delta}_j(t)$ through the new constructed varying coefficient model, i.e., $\boldsymbol{R}_{\cdot j}(t) = \mathbf{U}_{1:q}\boldsymbol{\delta}_j(t) + \tilde{\boldsymbol{\eta}}_{\cdot j}(t) + \tilde{\boldsymbol{\epsilon}}_{\cdot j}(t)$, where $\tilde{\boldsymbol{\eta}}_{\cdot j}$ and $\tilde{\boldsymbol{\epsilon}}_{\cdot j}$ are defined the same way as $\boldsymbol{\eta}_{\cdot j}$ and $\boldsymbol{\epsilon}_{\cdot j}$. For a fixed bandwidth h_δ, we can derive the estimator of $\boldsymbol{\delta}_j(t)$, as $\mathbf{U}_{1:q}^\top\sum_{k=1}^{n_v} a_k(h_\delta, t)\boldsymbol{R}_{\cdot j}(t_k)$, and the bias-corrected version, denoted as $\widehat{\boldsymbol{\delta}}_j(t)$. Then, we can construct the estimating equation:

$$\mathbf{X}\widetilde{\mathbf{B}}(t) + \mathbf{U}_{1:q}\widehat{\mathbf{D}}(t) = \mathbf{X}\mathbf{B}(t) + \mathbf{G}\widehat{\mathbf{D}}(t), \tag{8}$$

where $\widetilde{\mathbf{B}}(t) = (\tilde{\boldsymbol{\beta}}_1(t), \tilde{\boldsymbol{\beta}}_2(t))$ and $\widehat{\mathbf{D}}(t) = (\widehat{\boldsymbol{\delta}}_1(t), \widehat{\boldsymbol{\delta}}_2(t))$. In addition, assuming that the row vectors of $\mathbf{B}(t)$ and the row vectors of $\boldsymbol{\Lambda}(t)$ are orthogonal after mean centering, we can derive the estimator of $\mathbf{G}$ as $\widehat{\mathbf{G}} = \mathbf{U}_{1:q} + \mathbf{X}\int_0^1 \widetilde{\mathbf{B}}(t)\mathbf{P}\widehat{\mathbf{D}}^\top(t)dt\boldsymbol{\Omega}^{-1}$, where $\boldsymbol{\Omega} = \int_0^1 \widehat{\mathbf{D}}(t)\mathbf{P}\widehat{\mathbf{D}}^\top(t)dt$, $\mathbf{P} = \mathbf{I}_2 - \mathbf{1}_2(\mathbf{1}_2^\top\mathbf{1}_2)^{-1}\mathbf{1}_2^\top$, and $\mathbf{1}_2 = (1, 1)^\top$. Finally, the bias-corrected estimator of $\mathbf{B}(t)$ is given by

$$\widehat{\mathbf{B}}(t) = \widetilde{\mathbf{B}}(t) - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\widehat{\mathbf{G}}\widehat{\mathbf{D}}(t). \tag{9}$$

Remark 1. *First, to derive the estimated covariance function of $\boldsymbol{\eta}(s)$, we smooth the individual functions $\eta(s)$ via applying the method in [26] on the residuals. Second, to select the optimal bandwidth in $\widehat{\mathbf{B}}(s)$ and $\widehat{\mathbf{D}}(s)$, we use the leave-one-curve-out cross-validation, whereas for the optimal bandwidth in $\widehat{\boldsymbol{\eta}}(s)$, we use the generalized cross-validation score method [29,30]. Third, the rank q in Step (ii) is unknown in practice. According to the simulation and empirical studies in [28], we consider the eigenvalue difference method [31] in our real-data analysis.*

2.4. Hypothesis Testing

We consider the hypothesis for coefficients $\mathbf{B}(t)$ of Model (4) as

$$\mathbb{H}_0 : \mathbf{C}\mathrm{vec}(\mathbf{B}(t)) = \mathbf{0}\ \forall\ t \in [0, 1]\ v.s.\ \mathbb{H}_1 : \mathbf{C}\mathrm{vec}(\mathbf{B}(t)) \neq \mathbf{0} \text{ for some } t \in [0, 1], \tag{10}$$

where $\mathbf{C}$ is a $r \times 2p$ matrix of coefficients with rank r and $\text{vec}(\cdot)$ represents the vectorization function. The Wald global test statistic is then calculated as follows:

$$T_n = \int_0^1 T_n(t)dt, \text{ and } T_n(t) = \boldsymbol{\zeta}^\top(t)[\mathbf{C}\widehat{\boldsymbol{\Sigma}}_\eta(t,t) \otimes (\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top)\mathbf{C}^\top]^{-1}\boldsymbol{\zeta}(t), \tag{11}$$

where $\boldsymbol{\zeta}(t) = \mathbf{C}\text{vec}(\widehat{\mathbf{B}}(t))$, $\widehat{\mathbf{M}} = (\mathbf{I}_p, \mathbf{0}_{q\times q})(\widehat{\mathbf{W}}^\top\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^\top$, and $\widehat{\mathbf{W}} = (\mathbf{X}, \widehat{\mathbf{G}})$.

Instead of the complicated asymptotic distribution under $\mathbb{H}_1$, we consider a wild bootstrap method discussed in [28] to obtain the null distribution of global test statistic T_n, which consists of four steps:

1. Fit Model (4) on $\mathbf{X}$ and $\boldsymbol{\Psi}(t_k)_{k=1}^{n_v}$ under $\mathbb{H}_0$ and compute all the coefficients $\widehat{\mathbf{B}}(t)$, $\widehat{\mathbf{G}}$, $\widehat{\mathbf{D}}(t)$, $\widehat{\boldsymbol{\eta}}(t)$, $\widehat{\boldsymbol{\epsilon}}(t)$, and the global test statistic T_n.
2. Generate independent random vectors $\boldsymbol{\tau}^{(m)}$ and $\boldsymbol{\tau}^{(m)}(t_k)$ from standard normal distribution $N(\mathbf{0}, \mathbf{I}_n)$ for $k = 1, \ldots, n_v$ and generate

$$\boldsymbol{\Psi}^{(m)}(t_k) = \mathbf{X}\widehat{\mathbf{B}}(t_k) + \widehat{\mathbf{G}}\widehat{\mathbf{D}}(t_k) + diag(\boldsymbol{\tau}^{(m)})\widehat{\boldsymbol{\eta}}(t_k) + diag(\boldsymbol{\tau}^{(m)}(t_k))\widehat{\boldsymbol{\epsilon}}(t_k). \tag{12}$$

3. Based on $\{\boldsymbol{\Psi}^{(m)}(t_k)\}_{k=1}^{n_v}$ from previous step and $\mathbf{X}$, recompute $\widehat{\mathbf{B}}^{(m)}(t_k)$ and the global test statistic $T_n^{(m)}$.
4. Repeat Steps 2 and 3 M times to obtain $\{T_n^{(1)}, \ldots, T_n^{(M)}\}$ and calculate the p-value as $p = \sum_{m=1}^{M} \mathbf{1}(T_n^{(m)} > T_n)/M$.

Remark 2. *Given the elastic shape representations, the asymptotic properties of the estimated functions, including $\widehat{\mathbf{B}}(t)$ and $\widehat{\boldsymbol{\Lambda}}(t)$, the asymptotic distribution of the global test statistic T_n (11) under the null hypothesis, and its asymptotic power under local alternative hypotheses have been systematically investigated in our recent work [28]. Due to space limitations, we do not claim these theoretical results here. Readers who are interested in them can find the assumptions used to facilitate the technical details and the detailed proof in the Supplementary Material of [28].*

3. Case Study

3.1. Data Description and Processing

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu (accessed on 27 September 2023)). The ADNI was launched in 2003 by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a USD 60 million, 5-year public–private partnership. The primary goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials. The principal investigator of this initiative is Michael W. Weiner, MD, at the VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The goal was to recruit 800 subjects, but the initial study (ADNI-1) is followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org (accessed on 27 September 2023).

We consider $n = 707$ MRI scans from both normal controls and individuals with mild cognitive impairment (MCI) or AD in the ADNI-1 study. The scans, which were performed on a variety of 1.5 Tesla MRI scanners with protocols individualized for each scanner, include standard T1-weighted images obtained using volumetric three-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. To obtain the contour of planar CC, we use *FreeSurfer* [32] to process each T1-weighted MRI, including motion correction, non-parametric non-uniform intensity normalization, affine transform to the MNI305 atlas, intensity normalization, skull-stripping, and automatic subcortical segmentation. Some quality control procedures are performed on each output image data. Then, through package *CCSeg* [33], each T1-weighted MRI image and tissue segmentation result is used to extract the planar CC contour data on the midsagittal slice, which contains 100 landmarks (Figure 1a–b). Given the coordinate functions of landmarks (Figure 1c), we extract the aligned SRVF shape representation. The resulting aligned SRVFs $\boldsymbol{\psi}_i(t)$ is shown in Figure 1d.

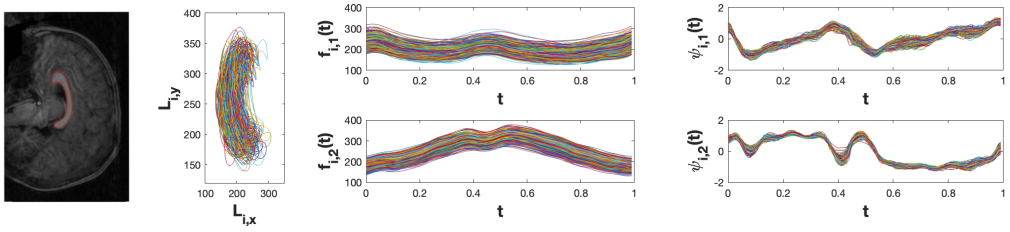

Figure 1. Preprocessing procedures for shape mediator: (**a**) raw MRI brain images with 2D CC segmented at the middle sagittal slice; (**b**) 50 landmarks sampled on the 2D contour of CC with $L_{i,x}$ and $L_{i,y}$ representing the x and y coordinate of CC respectively; (**c**) coordinate functions of landmarks with rotation, translation, and scaling removed; and (**d**) aligned SRVF representation of shape mediators.

3.2. Data Analysis

Research indicates that factors such as age, sex, handedness, and educational level can influence the structure of CC, which is linked to cognitive development [34–37]. Additionally, the APOE4 gene is not only linked to brain structural and functional variations across a wide age range, but is also recognized as a hereditary risk for declining cognitive ability and Alzheimer's disease [38,39]. Given the CC elastic shape responses, we are interested in their relationship with the demographic and clinical variables such as age, gender, handiness, education level, APOE-4, and one SNP variant, rs11719939 from chromosome 3, which is close to the AD-related high-risk gene, ATP2B2 [40,41]. The variable APOE-4 is coded as the number of alleles (zero, one, and two: zero and two represent the homozygous genotype, and one represents the heterozygous genotype). In addition, the population stratification is addressed by including the top two principal components (PCs) computed from the whole SNP data. Table 1 summarizes the demographic and genetic information of all the subjects.

Table 1. Demographic and genetic information about ADNI data: gender, range of age (RA), handiness, range of education level (REL), APOE-4, and SNP rs11719939.

Variable	Male	Female
Gender	420	287
RA (years)	[54.40, 89.30]	[55.10, 90.90]
Handiness (R/L)	386/34	266/21
REL (years)	[6, 20]	[6, 20]
APOE-4 (0/1/2)	205/169/46	147/107/33
SNP rs11719939 (0/1/2)	234/154/32	163/110/14

This study aimed to determine the effects of different covariates of interest on the CC shape alterations in the presence of unobserved confounding factors. We fitted our model with the real dataset, and estimated the regression coefficients and hidden factors. In particular, we used an eigenvalue difference method [31] to determine the number of hidden factors, q, in our model. We employed the wild bootstrap method and generated 500 bootstrap samples to derive the empirical null distribution of Wald's global test statistic and calculated the p-value for each regression coefficient function. For comparison, we considered two competing methods, i.e., the multivariate varying coefficient model (MVCM) developed in [30] and ComBat [2], which uses empirical Bayes estimates to remove the site effect.

According to the testing results in Table 2, our method detected significant effects related to all the covariates of interest, including age, gender, handiness, APOE-4, education level, two genetic PCs, and the causal SNP rs11719939, on the CC shapes. In particular, our model detected not only significant effects related to age, gender, and SNP, which are in agreement with the other two methods, but also the effects related to handiness, APOE-4, genetic PCs, and education, which were not fully detected at all by MVCM and ComBat.

Table 2. Hypothesis testing of estimated varying coefficients functions $\widehat{\beta}(s)$.

Variable	*p*-Value		
	Our Method [28]	**MVCM [30]**	**ComBat [2]**
Gender	0.022 *	0.000 *	0.025 *
Age	0.004 *	0.000 *	0.045 *
Handiness	0.016 *	0.880	0.466
APOE-4	0.008 *	0.348	0.364
PC1	0.002 *	0.540	0.340
PC2	0.002 *	0.008 *	0.225
Education	0.004 *	0.170	0.363
SNP rs11719939	0.014 *	0.000 *	0.000 *

*—significant at $\alpha = 0.05$.

In Table 3, we observed that gender, age, APOE-4, and SNP are significantly correlated with the first hidden factor, which indicates that the confounding factors are correlated with primary variables. This could be a potential cause of additional heterogeneity in elastic shape data. Therefore, this correlation must be accounted for while estimating the varying coefficient functions of the model.

Table 3. Correlation between hidden factors and primary variables.

Variable	Hidden Factors	
	Factor 1	**Factor 2**
Gender	−0.140	0.007
	(0.002)	(0.887)
Age	−0.113	−0.003
	(0.003)	(0.934)
Handiness	0.020	−0.001
	(0.772)	(0.989)
APOE4	−0.109	0.002
	(0.011)	(0.970)
PC1	−0.041	0.009
	(0.271)	(0.819)
PC2	0.048	0.0141
	(0.199)	(0.708)
Education	−0.049	−0.001
	(0.197)	(0.989)
SNP rs11719939	−0.209	0.011
	(0.000)	(0.794)

In the second part of our analysis, we considered the fivefold cross-validation strategy and computed the estimation error for the CC shape data using all three methods (Table 4). Our method outperformed both MVCM and ComBat in terms of both root-mean-squared error (RMSE) and mean absolute error (MAE), which indicates that our method successfully detected the potential hidden factors and captured the relationship between the elastic shape responses and covariates of interest better than the other two methods.

Table 4. Estimation error calculated for the three competing methods.

	MAE	**RMSE**
Our method [28]	0.001662	0.001752
MVCM [30]	0.003137	0.003269
ComBat [2]	0.003143	0.003276

4. Discussion

In this paper, we proposed a shape-on-scalar regression model that incorporates confounder adjustment. In particular, we leveraged the square root velocity function to extract elastic shape representations, which are embedded within the linear Hilbert space of square integrable functions. Subsequently, we introduced a shape regression model aimed at characterizing the intricate relationship between elastic shapes and covariates of interest, all while effectively managing the challenges posed by imaging heterogeneity. We developed comprehensive procedures for estimating and making inferences about the unknown model parameters. Through the real-data analysis, our method demonstrated its superiority in terms of estimation accuracy when compared to existing approaches.

In our case study on ADNI CC shape data, compared to other competing methods, i.e., MVCM and ComBat, our method identified a potential significant effect related to the education level, which is consistent with the existing literature [42], indicating that having ongoing learning experiences could be a important prevention strategy for cognitive impairment diseases such as AD. Our method also showed significant effects related to age, gender, and SNP on CC shape alterations.

Although our method demonstrated its success in the case study on ADNI CC shape data, there are couple of issues to be addressed. First, as discussed in [28], the key assumption of our method requires the row vectors of $\mathbf{B}(t)$ and the row vectors of $\mathbf{A}(t)$ to be orthogonal with respect to the underlying density function $p(t)$ after mean centering. Actually, this assumption is reasonable but difficult to examine in practice. Therefore, it is of great importance to improve the performance of our method even if this assumption does not hold. Second, this paper only investigated the linear relationship between the shape responses and other observed and/or hidden covariates. However, this assumption is not reliable when the nonlinear relationship exists in real applications. Therefore, we will find a scope to explore the nonlinear relationship between the covariates and the elastic shape response while adjusting for hidden factors.

Author Contributions: Conceptualization, C.H. and R.L.; methodology, C.H.; formal analysis, H.D.; investigation, H.D.; data curation, S.D.; writing—original draft preparation, H.D.; writing—review and editing, C.H.; visualization, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the US National Science Foundation Division of Mathematical Sciences grant DMS-1953087.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mueller, S.G.; Weiner, M.W.; Thal, L.J.; Petersen, R.C.; Jack, C.; Jagust, W.; Trojanowski, J.Q.; Toga, A.W.; Beckett, L. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N. Am.* **2005**, *15*, 869–877. [CrossRef] [PubMed]
2. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef]
3. Pomponio, R.; Erus, G.; Habes, M.; Doshi, J.; Srinivasan, D.; Mamourian, E.; Bashyam, V.; Nasrallah, I.M.; Satterthwaite, T.D.; Fan, Y.; et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage* **2020**, *208*, 116450. [CrossRef] [PubMed]
4. Leek, J.T.; Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **2007**, *3*, e161. [CrossRef] [PubMed]
5. Guan, H.; Liu, Y.; Yang, E.; Yap, P.T.; Shen, D.; Liu, M. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Med Image Anal.* **2021**, *71*, 102076. [CrossRef]
6. An, L.; Chen, J.; Chen, P.; Zhang, C.; He, T.; Chen, C.; Zhou, J.H.; Yeo, B.T.; Alzheimer's Disease Neuroimaging Initiative; Australian Imaging Biomarkers and Lifestyle Study of Aging. Goal-specific brain MRI harmonization. *Neuroimage* **2022**, *263*, 119570. [CrossRef]
7. Bayer, J.M.; Thompson, P.M.; Ching, C.R.; Liu, M.; Chen, A.; Panzenhagen, A.C.; Jahanshad, N.; Marquand, A.; Schmaal, L.; Sämann, P.G. Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. *Front. Neurol.* **2022**, *13*, 923988. [CrossRef]
8. Acquitter, C.; Piram, L.; Sabatini, U.; Gilhodes, J.; Moyal Cohen-Jonathan, E.; Ken, S.; Lemasson, B. Radiomics-based detection of radionecrosis using harmonized multiparametric MRI. *Cancers* **2022**, *14*, 286. [CrossRef]
9. Hu, F.; Chen, A.A.; Horng, H.; Bashyam, V.; Davatzikos, C.; Alexander-Bloch, A.; Li, M.; Shou, H.; Satterthwaite, T.D.; Yu, M.; et al. Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage* **2023**, *274*, 120125. [CrossRef]
10. Fortin, J.P.; Parker, D.; Tunç, B.; Watanabe, T.; Elliott, M.A.; Ruparel, K.; Roalf, D.R.; Satterthwaite, T.D.; Gur, R.C.; Gur, R.E.; et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **2017**, *161*, 149–170. [CrossRef]
11. Fortin, J.P.; Cullen, N.; Sheline, Y.I.; Taylor, W.D.; Aselcioglu, I.; Cook, P.A.; Adams, P.; Cooper, C.; Fava, M.; McGrath, P.J.; et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **2018**, *167*, 104–120. [CrossRef] [PubMed]
12. Jirsaraie, R.J.; Kaczkurkin, A.N.; Rush, S.; Piiwia, K.; Adebimpe, A.; Bassett, D.S.; Bourque, J.; Calkins, M.E.; Cieslak, M.; Ciric, R.; et al. Accelerated cortical thinning within structural brain networks is associated with irritability in youth. *Neuropsychopharmacology* **2019**, *44*, 2254–2262. [CrossRef] [PubMed]
13. Yu, M.; Linn, K.A.; Cook, P.A.; Phillips, M.L.; McInnis, M.; Fava, M.; Trivedi, M.H.; Weissman, M.M.; Shinohara, R.T.; Sheline, Y.I. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum. Brain Mapp.* **2018**, *39*, 4213–4227. [CrossRef] [PubMed]
14. Yamashita, A.; Yahata, N.; Itahashi, T.; Lisi, G.; Yamada, T.; Ichikawa, N.; Takamura, M.; Yoshihara, Y.; Kunimatsu, A.; Okada, N.; et al. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biol.* **2019**, *17*, e3000042. [CrossRef] [PubMed]
15. Bookstein, F.L. *Morphometric Tools for Landmark Data: Geometry and Biology*; Cambridge University Press: Cambridge, UK, 1991.
16. Small, C.G. *The Statistical Theory of Shape*; Springer: New York, NY, USA, 1996.
17. Kendall, D.G.; Barden, D.; Carne, T.K.; Le, H. *Shape and Shape Theory*; Wiley: Hoboken, NJ, USA, 1999.
18. Huang, C.; Srivastava, A.; Liu, R. Geo-FARM: Geodesic factor regression model for misaligned pre-shape responses in statistical shape analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; IEEE Computer Society: Washington, DC, USA, 2021; pp. 11496–11505.
19. Walterfang, M.; Luders, E.; Looi, J.C.; Rajagopalan, P.; Velakoulis, D.; Thompson, P.M.; Lindberg, O.; Östberg, P.; Nordin, L.E.; Svensson, L.; et al. Shape analysis of the corpus callosum in Alzheimer's disease and frontotemporal lobar degeneration subtypes. *J. Alzheimer's Dis.* **2014**, *40*, 897–906. [CrossRef]
20. Di Paola, M.; Spalletta, G.; Caltagirone, C. In vivo structural neuroanatomy of corpus callosum in Alzheimer's disease and mild cognitive impairment using different MRI techniques: A review. *J. Alzheimer's Dis.* **2010**, *20*, 67–95. [CrossRef]
21. Wang, X.D.; Ren, M.; Zhu, M.W.; Gao, W.P.; Zhang, J.; Shen, H.; Lin, Z.G.; Feng, H.L.; Zhao, C.J.; Gao, K. Corpus callosum atrophy associated with the degree of cognitive decline in patients with Alzheimer's dementia or mild cognitive impairment: A meta-analysis of the region of interest structural imaging studies. *J. Psychiatr. Res.* **2015**, *63*, 10–19. [CrossRef]
22. Jiang, Z.; Yang, H.; Tang, X. Deformation-based statistical shape analysis of the corpus callosum in mild cognitive impairment and Alzheimer's disease. *Curr. Alzheimer Res.* **2018**, *15*, 1151–1160. [CrossRef]
23. Kamal, S.; Park, I.; Kim, Y.J.; Kim, Y.J.; Lee, U. Alteration of the corpus callosum in patients with Alzheimer's disease: Deep learning-based assessment. *PLoS ONE* **2021**, *16*, e0259051. [CrossRef]
24. Srivastava, A.; Klassen, E. *Functional and Shape Data Analysis*; Springer: New York, NY, USA, 2016.
25. Srivastava, A.; Wu, W.; Kurtek, S.; Klassen, E.; Marron, J.S. Registration of functional data using Fisher-Rao metric. *arXiv* **2011**, arXiv:1103.3817.
26. Ruppert, D.; Wand, M.P. Multivariate locally weighted least squares regression. *Ann. Stat.* **1994**, *22*, 1346–1370. [CrossRef]

27. Fan, J.; Gijbels, I. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*; CRC Press: Chapman Hall, London, 1996; Volume 66.
28. Huang, C.; Zhu, H. Functional hybrid factor regression model for handling heterogeneity in imaging studies. *Biometrika* **2022**, *109*, 1133–1148. [CrossRef] [PubMed]
29. Zhang, J.; Chen, J. Statistical inference for functional data. *Ann. Stat.* **2007**, *35*, 1052–1079. [CrossRef]
30. Zhu, H.; Li, R.; Kong, L. Multivariate varying coefficient model for functional responses. *Ann. Stat.* **2012**, *40*, 2634–2666. [CrossRef]
31. Onatski, A. Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.* **2010**, *92*, 1004–1016. [CrossRef]
32. Fischl, B. FreeSurfer. *Neuroimage* **2012**, *62*, 774–781. [CrossRef] [PubMed]
33. Vachet, C.; Yvernault, B.; Bhatt, K.; Smith, R.G.; Gerig, G.; Hazlett, H.C.; Styner, M. Automatic corpus callosum segmentation using a deformable active Fourier contour model. In Proceedings of the Medical Imaging 2012: Biomedical Applications in Molecular, Structural, and Functional Imaging, San Diego, CA, USA, 4–9 February 2012; Volume 8317, pp. 79–85.
34. Prendergast, D.M.; Ardekani, B.; Ikuta, T.; John, M.; Peters, B.; DeRosse, P.; Wellington, R.; Malhotra, A.K.; Szeszko, P.R. Age and sex effects on corpus callosum morphology across the lifespan. *Hum. Brain Mapp.* **2015**, *36*, 2691–2702. [CrossRef]
35. Rushton, J.P.; Ankney, C.D. Brain size and cognitive ability: Correlations with age, sex, social class, and race. *Psychon. Bull. Rev.* **1996**, *3*, 21–36. [CrossRef]
36. Deary, I.J.; Corley, J.; Gow, A.J.; Harris, S.E.; Houlihan, L.M.; Marioni, R.E.; Penke, L.; Rafnsson, S.B.; Starr, J.M. Age-associated cognitive decline. *Br. Med. Bull.* **2009**, *92*, 135–152. [CrossRef]
37. Guadalupe, T.; Willems, R.M.; Zwiers, M.P.; Arias Vasquez, A.; Hoogman, M.; Hagoort, P.; Fernandez, G.; Buitelaar, J.; Franke, B.; Fisher, S.E.; et al. Differences in cerebral cortical anatomy of left-and right-handers. *Front. Psychol.* **2014**, *5*, 261. [CrossRef]
38. Matura, S.; Prvulovic, D.; Jurcoane, A.; Hartmann, D.; Miller, J.; Scheibe, M.; O'Dwyer, L.; Oertel-Knöchel, V.; Knöchel, C.; Reinke, B.; et al. Differential effects of the ApoE4 genotype on brain structure and function. *Neuroimage* **2014**, *89*, 81–91. [CrossRef] [PubMed]
39. Montagne, A.; Nation, D.A.; Sagare, A.P.; Barisano, G.; Sweeney, M.D.; Chakhoyan, A.; Pachicano, M.; Joe, E.; Nelson, A.R.; D'Orazio, L.M.; et al. APOE4 leads to blood–brain barrier dysfunction predicting cognitive decline. *Nature* **2020**, *581*, 71–76. [CrossRef] [PubMed]
40. Berrocal, M.; Marcos, D.; Sepúlveda, M.R.; Pérez, M.; Ávila, J.; Mata, A.M. Altered Ca^{2+} dependence of synaptosomal plasma membrane Ca^{2+}-ATPase in human brain affected by Alzheimer's disease. *FASEB J.* **2009**, *23*, 1826–1834. [CrossRef] [PubMed]
41. Berridge, M.J. Calcium signalling remodelling and disease. *Biochem. Soc. Trans.* **2012**, *40*, 297–309. [CrossRef] [PubMed]
42. Castro-Caldas, A.; Miranda, P.C.; Carmo, I.; Reis, A.; Leote, F.; Ribeiro, C.; Ducla-Soares, E. Influence of learning to read and write on the morphology of the corpus callosum. *Eur. J. Neurol.* **1999**, *6*, 23–28. [CrossRef] [PubMed]