Intrinsic Regression Models for Medial Representation of Subcortical Structures *

Xiaoyan Shi, Hongtu Zhu, Joseph G. Ibrahim,

Faming Liang, Jeffrey Lieberman, and Martin Styner

*Address for correspondence and reprints: Hongtu Zhu, Ph.D., hzhu@bios.unc.edu. Department of Biostatistics, Gillings School of Global Public Health, 3109 McGavran-Greenberg Hall, Campus Box 7420, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420. H. Zhu is Professor of Biostatistics (E-mail: hzhu@bios.unc.edu), J. G. Ibrahim is Alumni Distinguished Professor of Biostatistics (E-mail: ibrahim@bios.unc.edu), and X. Shi was Ph.d student (E-mail: amy.shi@sas.com), Department of Biostatistics and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, NC 27599-7420. F. Liang is Professor of Statistics (E-mail: fliang@stat.tamu.edu), Department of Statistics, Texas A & M University, College Station, TX 77843-3143. Jeffrey Lieberman is Lawrence C. Kolb Professor of Psychiatry (E-mail: jlieberman@pi.cpmc.Columbia.edu), Department of Psychiatry, Columbia University Medical Center, 1051 Riverside Drive, New York, New York 10032, U.S.A. M. Styner is Assistant Professor (E-mail: yasheng.chen@med.unc.edu), Department of Computer Science and Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. This work was supported in part by NIH grants UL1-RR025747-01, R21AG033387, P01CA142538-01, MH086633, GM 70335, and CA 74015 to Drs. Zhu and Ibrahim, DMS-1007457 and DMS-1106494 to Dr. Liang, and Lilly Research Laboratories, the UNC NDRC HD 03110, Eli Lilly grant F1D-MC-X252, and NIH Roadmap Grant U54 EB005149-01, NAMIC to Dr. Styner. We thank the Editor, an associated editor, and two references for help suggestions, which have improved the present form of this article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or the NIH.

Abstract

The aim of this paper is to develop a semiparametric model for describing the variability of the medial representation of subcortical structures, which belongs to a Riemannian manifold, and establishing its association with covariates of interest, such as diagnostic status, age and gender. We develop a two-stage estimation procedure to calculate the parameter estimates. The first stage is to calculate an intrinsic least squares estimator of the parameter vector using the annealing evolutionary stochastic approximation Monte Carlo algorithm and then the second stage is to construct a set of estimating equations to obtain a more efficient estimate with the intrinsic least squares estimate as the starting point. We use Wald statistics to test linear hypotheses of unknown parameters and establish their limiting distributions. Simulation studies are used to evaluate the accuracy of our parameter estimates and the finite sample performance of the Wald statistics. We apply our methods to the detection of the difference in the morphological changes of the left and right hippocampi between schizophrenia patients and healthy controls using medial shape description.

Keywords: Intrinsic least squares estimator; Medial representation; Semiparametric model; Wald statistic.

1 Introduction

The medial representation of subcortical structures provides a useful framework for describing shape variability in local thickness, bending, and widening for subcortical structures (Fletcher et al., 2004). In the medial representation framework, a geometric object is represented as a set of connected continuous medial primitives, called medial atoms. See Figure 1 for a hippocampus example. For 3-dimensional objects, these medial atoms are formed by the centers of the inscribed spheres and by the associated spokes from the sphere centers to the two respective tangent points on the object boundary. Specifically, a medial atom $\mathbf{m} = (\mathbf{O}^T, r, \mathbf{s}_0^T, \mathbf{s}_1^T)^T$ is formed by a position $\mathbf{O}$, the center of the inscribed sphere; a radius r, the common spoke length; and $(\mathbf{s}_0, \mathbf{s}_1)$, the two unit spoke directions (Pizer et al., 2003; Styner et al., 2004). A medial atom can be regarded as a point on a Riemannian manifold, $M(1) = R^3 \times R^+ \times S^2 \times S^2$, where S^2 is the sphere in R^3 with radius one. A medial representation model consisting of K medial atoms can be described as the direct product of K copies of $M(1)$, i.e., $M(1)^K = \prod_{i=1}^K M(1)$. The existing statistical analytical methods for the medial representation include principal geodesic analysis, the estimation of extrinsic and intrinsic means, and a permutation test for comparing medial representation data from two groups (Fletcher et al., 2004). The scientific interests of some neuroimaging studies, however, typically focus on establishing the association between subcortical structure and a set of covariates, particularly diagnostic status, age, and gender, thus requiring a regression modeling framework for medial representation.

There are several challenging issues including multiple directions on S^2 and the complex correlation structure among different components of $M(1)$ in developing medial representation regression models with a set of covariates. Although there is a sparse literature on regression modeling of a single directional response and a set of covariates of interest (Mardia and Jupp, 1983; Jupp and Mardia 1989), these regression models of directional data are based on particular parametric distributions, such as the von Mises-Fisher distribution (Mardia, 1975; Mardia and

Jupp, 1983; Presnell et al., 1998). For instance, existing circular regression models assume that the angular response follows the von Mises-Fisher distribution with either the angular mean η_i or the concentration parameter κ_i being associated with the covariates $\mathbf{x}_i$ (Gould, 1969; Johnson and Wehrly, 1978; Fisher and Lee, 1992). However, it remains unknown whether it is appropriate to directly apply these parametric models for a single directional measure to simultaneously characterize the two spoke directions at each atom, which are correlated. Moreover, the two spoke directions may be correlated with other components of each atom and this provides further challenges in developing a parametric model to simultaneously model all components of each atom of the medial representation.

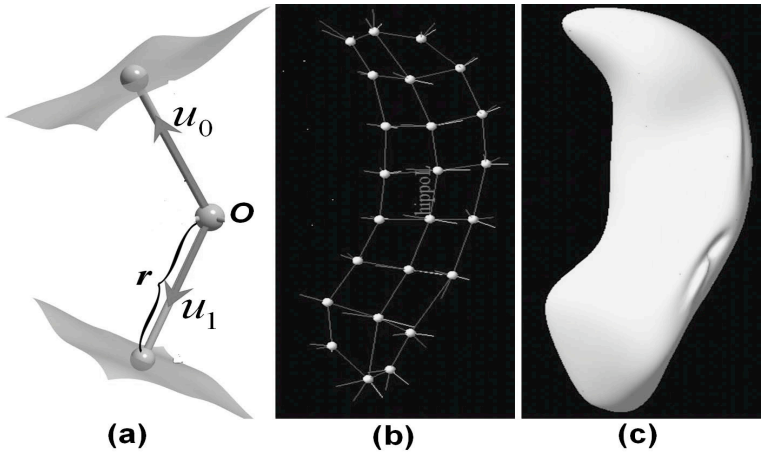


Figure 1: (a) A medial representation model $\mathbf{m} = (\mathbf{O}^T, r, \mathbf{s}_0^T, \mathbf{s}_1^T)^T$ at an atom, where $\mathbf{O}$ is the center of the inscribed sphere, r is the common spoke length, and $(\mathbf{s}_0, \mathbf{s}_1)$ are the two unit spoke directions; (b) a skeleton of a hippocampus with 24 medial atoms; (c) the smoothed surface of the hippocampus.

The rest of this paper is organized as follows. In Section 2, we formulate the semiparametric regression model and introduce the two-stage estimation procedure for estimating the regression coefficients. Then, we establish asymptotic properties of our estimates and then develop Wald statistics to carry out hypothesis testing. Simulation studies in Section 3 are used to assess the finite sample performance of the parameter estimates and Wald test statistics. In Section

4, we illustrate the application of our statistical methods to the detection of the difference in morphological changes of the hippocampi between schizophrenia patients and healthy controls in a neuroimaging study of schizophrenia.

2 Theory

2.1 Inverse Link functions

Suppose we have an exogenous $q \times 1$ covariate vector $\mathbf{x}_i$ and a medial representation for a particular sub-cortical structure, denoted by $M_i = \{\mathbf{m}_i(d) : d \in \mathcal{D}\}$, for the $i-$th subject, where d represents an atom of the medial representation. For notational simplicity, we temporarily drop atom d from our notation. We formally introduce a semiparametric regression model for medial representation responses and covariates of interest from n subjects. The regression model involves modeling a conditional mean of a medial representation response $\mathbf{m}_i$ at an atom given $\mathbf{x}_i$, denoted by $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{\mu}(\mathbf{x}_i, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients in $\mathcal{B} \subset R^p$. Thus, $\boldsymbol{\mu}(\cdot, \cdot)$ is a map from $R^q \times R^p$ to $M(1)$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\boldsymbol{\mu}_{oi}(\boldsymbol{\beta})^T, \mu_{ri}(\boldsymbol{\beta}), \boldsymbol{\mu}_{0i}(\boldsymbol{\beta})^T, \boldsymbol{\mu}_{1i}(\boldsymbol{\beta})^T)^T$, which is a 10×1 vector and $\boldsymbol{\mu}_{oi}(\boldsymbol{\beta})$, $\mu_{ri}(\boldsymbol{\beta})$, $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$, and $\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})$ are the 'conditional means' of the location $\mathbf{O}_i$, the radius r_i, and the two spoke directions $\mathbf{s}_{0i}$ and $\mathbf{s}_{1i}$ respectively, given $\mathbf{x}_i$, for the i-th subject. Note that for spoke directions, we borrow the term conditional mean for random variables in Euclidean space.

We need to formalize the notion of conditional mean explicitly. For the location component of a medial representation, we may set $\boldsymbol{\mu}_{oi}(\boldsymbol{\beta}) = (g_1(\mathbf{x}_i, \boldsymbol{\beta}_1), g_2(\mathbf{x}_i, \boldsymbol{\beta}_2), g_3(\mathbf{x}_i, \boldsymbol{\beta}_3))^T$, where $g_k(\cdot, \cdot)$ is a known inverse link function and $\boldsymbol{\beta}_k$ is a $p_k \times 1$ coefficient vector for $k = 1, 2, 3$. There are many different ways of specifying $g_k(\mathbf{x}_i, \boldsymbol{\beta}_k)$. The simplest one is the linear inverse link function $g_k(\mathbf{x}_i, \boldsymbol{\beta}_k) = \mathbf{x}_i^T \boldsymbol{\beta}_k$. We may also represent $g_k(\mathbf{x}_i, \boldsymbol{\beta}_k)$ as a linear combination of basis functions $\{\psi_j(\mathbf{x}_i) : j = 1, \ldots, J\}$, such as B-splines, that is $g_k(\mathbf{x}_i, \boldsymbol{\beta}_k) = \sum_{j=1}^{J} \psi_j(\mathbf{x}_i)\beta_{kj}$, in which β_{kj} is the j-th component of $\boldsymbol{\beta}_k$. In this way, we can approximate a nonlinear function of $\mathbf{x}_i$ using the

linear combination of basis functions. For the radius component, we may use $\mu_{ri}(\boldsymbol{\beta}) = g_4(\mathbf{x}_i, \boldsymbol{\beta}_4)$, where $\boldsymbol{\beta}_4$ is a $p_4 \times 1$ coefficient vector for a medial representation radius. Since a radius is always positive, a natural inverse link function is $g_4(\mathbf{x}_i, \boldsymbol{\beta}_4) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_4)$, among other possible choices.

As the two spoke directions at each atom of a medial representation are spherical responses, we develop a link function $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}) \in S^2$ for the first spoke direction at a specific atom for notational simplicity. Let $\mathbf{x}_{i,d}$ be a $q_d \times 1$ vector of all the discrete covariates, $\mathbf{x}_{i,c}$ are a $q_c \times 1$ vector of all the continuous covariates and their potential interactions with $\mathbf{x}_{i,d}$, $\boldsymbol{\beta}_{5d}$ and $\boldsymbol{\beta}_{5c}$ are the regression parameters corresponding to $\mathbf{x}_{i,d}$ and $\mathbf{x}_{i,c}$, respectively, and $\boldsymbol{\beta}_5$ contains all unknown parameters in $\boldsymbol{\beta}_{5d}$ and $\boldsymbol{\beta}_{5c}$. From now on, all covariates have been centered to have mean zero. We assume that all first spoke directions associated with the same discrete covariate vector $\mathbf{x}_{i,d}$ are concentrated around a center on the sphere given by

$$\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d}) = (\sin(\theta(\mathbf{x}_{i,d}))\cos(\phi(\mathbf{x}_{i,d})), \sin(\theta(\mathbf{x}_{i,d}))\sin(\phi(\mathbf{x}_{i,d})), \cos(\theta(\mathbf{x}_{i,d})))^T, \tag{1}$$

where $\theta(\mathbf{x}_{i,d})$ and $\phi(\mathbf{x}_{i,d})$ are, respectively, the colatitude and the longitude, and $\boldsymbol{\beta}_{5d}$ includes all unknown parameters $\theta(\mathbf{x}_{i,d})$ and $\phi(\mathbf{x}_{i,d})$ for different $\mathbf{x}_{i,d}$.

We then describe the stereographic projection of projecting $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ on the plane with base point $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})$, denoted by $T_{st;\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}(\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}))$ (Downs, 2003). A graphic illustration of the stereographic projection $T^{-1}_{st;(0,0,1)}(u, v, -1)$ is given in Figure 2 (a). The stereographic projection $T_{st;\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}(\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}))$ is defined as the point of intersection for the plane passing through $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})$ with the normal vector $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})$, which is given by $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})^T\{(u, v, w)^T - \mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})\} = 0$ for $(u, v, w) \in R^3$, and the line passing through $-\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})$ and $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$: $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}) - t\{\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d}) + \boldsymbol{\mu}_{0i}(\boldsymbol{\beta})\}$ for $t \in (-\infty, \infty)$. With some calculation, it can be shown that $T_{st;\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}(\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}))$ is given by

$$T_{st;\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}(\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})) = \frac{2\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})}{1 + \boldsymbol{\mu}_{0i}(\boldsymbol{\beta})^T\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})} - \frac{\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})\{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})^T\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d}) - 1\}}{1 + \boldsymbol{\mu}_{0i}(\boldsymbol{\beta})^T\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})}.$$

Let $\mathbf{R}$ be a rotation matrix in SO(3) such that $\mathbf{R}^T = \mathbf{R}^{-1}$ and $\det(\mathbf{R}) = 1$, where $\det(\mathbf{R})$ denotes the determinant of $\mathbf{R}$ and SO(3) is the set of 3×3 rotation matrices. By applying the rotation

matrix $\mathbf{R}$ to both $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})$ and $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$, we have

$$T_{st;R\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}(\mathbf{R}\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})) = \mathbf{R}T_{st;\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}(\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})). \tag{2}$$

We consider a specific rotation matrix for rotating $\mathbf{s}_1 = (s_{1,u}, s_{1,v}, s_{1,w})^T \in S^2$ to $\mathbf{s}_2 = (s_{2,u}, s_{2,v}, s_{2,w})^T \in S^2$, denoted by $\mathbf{R}_{s_1,s_2}$, such that $\mathbf{R}_{s_1,s_2}\mathbf{s}_1 = \mathbf{s}_2$. We need to calculate $\eta = \arccos(\mathbf{s}_1^T\mathbf{s}_2) = \arccos(s_{1,u}s_{2,u} + s_{1,v}s_{2,v} + s_{1,w}s_{2,w})$ and $\mathbf{s}_3 = \mathbf{s}_1 \times \mathbf{s}_2 / \|\mathbf{s}_1 \times \mathbf{s}_2\| = (s_{3,u}, s_{3,v}, s_{3,w})^T$, where $\mathbf{s}_1 \times \mathbf{s}_2 = (s_{1,v}s_{2,w} - s_{1,w}s_{2,v}, s_{1,w}s_{2,u} - s_{1,u}s_{2,w}, s_{1,u}s_{2,v} - s_{1,v}s_{2,u})^T$ and $\|\cdot\|$ is the Euclidean norm of a vector. Then, $\mathbf{R}_{s_1,s_2}$ is given by

$$\begin{pmatrix} s_{3,u}^2 c_\eta + \cos(\eta), & s_{3,u}s_{3,v}c_\eta - s_{3,w}\sin(\eta), & s_{3,u}s_{3,w}c_\eta + s_{3,v}\sin(\eta) \\ s_{3,u}s_{3,v}c_\eta + s_{3,w}\sin(\eta), & s_{3,v}^2 c_\eta + \cos(\eta), & s_{3,v}s_{3,w}c_\eta - s_{3,u}\sin(\eta) \\ s_{3,u}s_{3,w}c_\eta - s_{3,v}\sin(\eta), & s_{3,v}s_{3,w}c_\eta + s_{3,u}\sin(\eta), & s_{3,w}^2 c_\eta + \cos(\eta) \end{pmatrix}, \tag{3}$$

where $c_\eta = 1 - \cos(\eta)$.

The inverse link function $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ is explicitly given as follows. By letting $\mathbf{R} = \mathbf{R}_{\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d}),(0,0,-1)^T}$ in (2), in which $(0,0,-1)^T$ is the south pole of S^2, we have

$$T_{st;(0,0,-1)^T}(\mathbf{R}_{\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d}),(0,0,-1)^T}\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})) = \mathbf{R}_{\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d}),(0,0,-1)^T}T_{st;\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}(\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})). \tag{4}$$

We assume that

$$T_{st;(0,0,-1)^T}(\mathbf{R}_{\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d}),(0,0,-1)^T}\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})) = (\mathbf{x}_{ic}^T\boldsymbol{\beta}_{5c}, -1)^T, \tag{5}$$

where $\boldsymbol{\beta}_{5c}$ is a $q_c \times 2$ matrix. Let $T^{-1}_{st;(0,0,-1)^T}$ be the inverse map of the stereographic projection mapping from the plane with base point $(0,0,-1)$ back to S^2 such that

$$T^{-1}_{st;(0,0,-1)^T}((u,v,-1)) = \left(\frac{4u}{u^2+v^2+4}, \frac{4v}{u^2+v^2+4}, \frac{u^2+v^2-4}{u^2+v^2+4}\right).$$

Please see Fig. 2 (a) for details. Note that $\mathbf{R}_{\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d}),(0,0,-1)^T} \in \mathrm{SO}(3)$, the inverse link function $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ is given by

$$\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}) = \mathbf{R}_{(0,0,-1)^T,\mathbf{g}_5(\mathbf{x}_{i,d},\beta_{5d})}T^{-1}_{st;(0,0,-1)^T}((\mathbf{x}_{i,c}^T\boldsymbol{\beta}_{5c}, -1)^T). \tag{6}$$

When $\boldsymbol{\beta}_{5c} = 0$ indicating no continuous covariate effect, $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ reduces to $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})$. Similarly, for the second spoke direction, we introduce $\boldsymbol{\beta}_{6d}$ and $\boldsymbol{\beta}_{6c}$ as the regression parameters corresponding to $\mathbf{x}_{i,d}$ and $\mathbf{x}_{i,c}$, respectively, and then we define $\mathbf{g}_6(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{6d})$ and $\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})$, respectively, as the center associated with the same discrete covariate vector $\mathbf{x}_{i,d}$ and the inverse link function by following (1) and (6). We have discussed various inverse link functions for $\boldsymbol{\mu}(\mathbf{x}_i, \boldsymbol{\beta})$, but these link functions can be misspecified for a given data set. To avoid such misspecification, we may estimate these inverse link functions nonparametrically. It is a topic for future research.

2.2 Intrinsic regression model

Now, we introduce a definition of a residual to ensure that $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ is the proper conditional mean of $\mathbf{m}_i$ given $\mathbf{x}_i$. For instance, in a classical linear model, the response is the sum of the regression function and the residual, and the conditional mean of the response equals the regression function. Given two points $\mathbf{m}_i$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ on the manifold, we need to define the residual or difference between them. At $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, we have the tangent space of $M(1)$, denoted by $T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$, which is a Euclidean space representing a first order approximation of the manifold $M(1)$ near $\boldsymbol{\mu}_i(\boldsymbol{\beta})$. We calculate the projection of $\mathbf{m}_i$ onto $T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$, denoted by $\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)$, as follows:

$$\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) = (\mathbf{O}_i - \boldsymbol{\mu}_{oi}(\boldsymbol{\beta}), \log(r_i/\mu_{ri}(\boldsymbol{\beta})), \mathrm{L}_{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})}(\mathbf{s}_{0i})^T, \mathrm{L}_{\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})}(\mathbf{s}_{1i})^T)^T, \tag{7}$$

where $\mathrm{L}_{\boldsymbol{\mu}_{ki}(\boldsymbol{\beta})}(s_{ki}) = \arccos(\boldsymbol{\mu}_{ki}(\boldsymbol{\beta})^T\mathbf{s}_{ki})\tilde{\mathbf{s}}_{ki}/||\tilde{\mathbf{s}}_{ki}||$, in which $\tilde{\mathbf{s}}_{ki} = \mathbf{s}_{ki} - \{\boldsymbol{\mu}_{ki}(\boldsymbol{\beta})^T\mathbf{s}_{ki}\}\boldsymbol{\mu}_{ki}(\boldsymbol{\beta})$ for $k = 0, 1$. Thus, $\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)$ can be regarded as the residual or difference between $\mathbf{m}_i$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ in $T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$. Geometrically, $\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)$ is associated with the Riemannian Exponential and Logarithm maps on $M(1)$.

We introduce the Riemannian Exponential and Logarithm maps on $M(1)$. Let the tangent vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_o, \theta_r, \boldsymbol{\theta}_{s_0}, \boldsymbol{\theta}_{s_1})^T \in T_mM(1)$, where $\boldsymbol{\theta}_o \in R^3$ is the location tangent component, $\theta_r \in R$ is the radius tangent component, and $\boldsymbol{\theta}_{s_0}$ and $\boldsymbol{\theta}_{s_1} \in R^3$ are the two directional tangent

components. Let $\gamma_{\mathbf{m}}(t;\boldsymbol{\theta})$ be the geodesic on $M(1)$ passing through $\gamma_{\mathbf{m}}(0;\boldsymbol{\theta}) = \mathbf{m} \in M(1)$ in the direction of the tangent vector $\boldsymbol{\theta} \in T_{\mathbf{m}}M(1)$. The Riemannian Exponential map, denoted by $\mathrm{Exp}_{\mathbf{m}}(\cdot)$, maps the tangent vector $\boldsymbol{\theta}$ at $\mathbf{m}$ to a point $\mathbf{m}_1 \in M(1)$ and $\mathrm{Exp}_{\mathbf{m}}(\boldsymbol{\theta}) = \gamma_{\mathbf{m}}(1;\boldsymbol{\theta})$. The Riemannian Logarithm map, denoted by $\mathrm{L}_{\mathbf{m}}(\mathbf{m}_1)$, maps $\mathbf{m}_1 \in M(1)$ onto the tangent vector $\boldsymbol{\theta} = \mathrm{L}_{\mathbf{m}}(\mathbf{m}_1) \in T_{\mathbf{m}}M(1)$. The Riemannian Exponential map and Logarithm map are inverses of each other, that is $\mathrm{Exp}_{\mathbf{m}}(\mathrm{L}_{\mathbf{m}}(\mathbf{m}_1)) = \mathbf{m}_1$.

Because a medial representation is the product space of several spaces, the Riemannian Exponential/Logarithm map for $M(1)$ is the product of the Riemannian Exponential/Logarithm maps for each space. Let $\mathbf{m} = (\mathbf{O}^T, r, \mathbf{s}_0^T, \mathbf{s}_1^T)^T$ and $\mathbf{m}_1 = (\mathbf{O}_1^T, r_1, \mathbf{s}_{0,1}^T, \mathbf{s}_{1,1}^T)^T$ be two points in $M(1)$ and $\boldsymbol{\theta} \in T_{\mathbf{m}}M(1)$. We give the explicit form of the Exponential and Logarithm maps for each space of interest. For the space of locations, $\mathrm{Exp}_o(\boldsymbol{\theta}_o) = \mathbf{O} + \boldsymbol{\theta}_o$, and $\mathrm{L}_o(\mathbf{O}_1) = \mathbf{O}_1 - \mathbf{O}$. For the space of radiuses, $\mathrm{Exp}_r(\theta_r) = r\exp(\theta_r)$ and $\mathrm{L}_r(r_1) = \log(r_1/r)$. For the space S^2, $\mathrm{Exp}_{\mathbf{s}_0}(\boldsymbol{\theta}_{s_0}) = \cos(\|\boldsymbol{\theta}_{s_0}\|_2)\mathbf{s}_0 + \sin(\|\boldsymbol{\theta}_{s_0}\|_2)\boldsymbol{\theta}_{s_0}/\|\boldsymbol{\theta}_{s_0}\|_2$. Let $\tilde{\mathbf{s}}_{0,1} = \mathbf{s}_{0,1} - (\mathbf{s}_0^T\mathbf{s}_{0,1})\mathbf{s}_0 \neq 0$. If $\mathbf{s}_0$ and $\mathbf{s}_{0,1}$ are not antipodal ($\mathbf{s}_0 \neq -\mathbf{s}_{0,1}$), we can get $\mathrm{L}_{\mathbf{s}_0}(\mathbf{s}_{0,1}) = \arccos(\mathbf{s}_0^T\mathbf{s}_{0,1})\tilde{\mathbf{s}}_{0,1}/\|\tilde{\mathbf{s}}_{0,1}\|_2$. Thus, for the space $M(1)$, the Riemannian Exponential and Logarithm maps are, respectively, given by

$$\mathrm{Exp}_{\mathbf{m}}(\boldsymbol{\theta}) = (\mathbf{O}^T + \boldsymbol{\theta}_o^T, r\exp(\theta_r), \mathrm{Exp}_{s_0}(\boldsymbol{\theta}_{s_0})^T, \mathrm{Exp}_{s_1}(\boldsymbol{\theta}_{s_1})^T)^T, \tag{8}$$

$$\mathrm{L}_{\mathbf{m}}(\mathbf{m}_1) = (\mathbf{O}_1^T - \mathbf{O}^T, \log(r_1/r), \mathrm{L}_{\mathbf{s}_0}(\mathbf{s}_{0,1})^T, \mathrm{L}_{\mathbf{s}_1}(\mathbf{s}_{1,1})^T)^T. \tag{9}$$

Although the $\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) \in T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$ are in different tangent spaces, we can use parallel transport to translate them to the same tangent space at an overall base point, denoted by $\mathbf{B}(\boldsymbol{\beta})$. We choose $\mathbf{B}(\boldsymbol{\beta}) = (0,0,0,1,\overline{\mathbf{g}}_5(\boldsymbol{\beta}_{5d})^T, \overline{\mathbf{g}}_6(\boldsymbol{\beta}_{6d})^T)^T$, where $\overline{\mathbf{g}}_5(\boldsymbol{\beta}_{5d})$ and $\overline{\mathbf{g}}_6(\boldsymbol{\beta}_{6d})$ are the mean directions of $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d})$ and $\mathbf{g}_6(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{6d})$ for all possible $\mathbf{x}_{i,d}$, respectively. We use parallel transport formulated by a rotation matrix,

$$\mathbf{R}(\boldsymbol{\mu}_i(\boldsymbol{\beta}) \Rightarrow \mathbf{B}(\boldsymbol{\beta})) = \mathrm{diag}\{\mathbf{I}_3, 1, \mathbf{R}_{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}), \overline{\mathbf{g}}_5(\boldsymbol{\beta}_{5d})}, \mathbf{R}_{\boldsymbol{\mu}_{1i}(\boldsymbol{\beta}), \overline{\mathbf{g}}_6(\boldsymbol{\beta}_{6d})}\}, \tag{10}$$

to translate $\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) \in T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$ into $\{\mathbf{R}(\boldsymbol{\mu}_i(\boldsymbol{\beta}) \Rightarrow \mathbf{B}(\boldsymbol{\beta}))\}\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) \in T_{\mathbf{B}(\boldsymbol{\beta})}M(1)$. An illustration of the parallel transport is given in Figure 2 (b). Finally, we define the rotated

residual of $\mathbf{m}_i$ with respect to $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ as

$$\mathcal{E}_i(\boldsymbol{\beta}) = \{\mathbf{R}(\boldsymbol{\mu}_i(\boldsymbol{\beta}) \Rightarrow \mathbf{B}(\boldsymbol{\beta}))\}\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) \quad \text{for} \quad i = 1, \ldots, n. \tag{11}$$

The $\mathcal{E}_i(\boldsymbol{\beta})$ are uniquely defined in the same tangent space $T_{\mathbf{B}(\boldsymbol{\beta})}M(1)$, which is a Euclidean space.

The intrinsic regression model for medial representations $M(1)$ at an atom is then defined by

$$E\{\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\} = \mathbf{0}, \qquad E[\{\mathbf{R}(\boldsymbol{\mu}_i(\boldsymbol{\beta}) \Rightarrow \mathbf{B}(\boldsymbol{\beta}))\}\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) \mid \mathbf{x}_i] = \mathbf{0} \tag{12}$$

for $i = 1, \ldots, n$, where the expectation is taken with respect to the conditional distribution of $\mathcal{E}_i(\boldsymbol{\beta})$ given $\mathbf{x}_i$ (Le, 2001). In model (12), the nonparametric component is the distribution of $\mathbf{m}_i$ given $\mathbf{x}_i$, which is left unspecified, while the parametric component is the mean function $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, which is assumed to be known. Moreover, our model (12) does not assume a homogeneous variance across all atoms and subjects. This is also desirable for real applications, because between-subject and between-atom variabilities can be substantial.

At atom d, let $\mathcal{E}_i(\boldsymbol{\beta}, d)$ be $\{\mathbf{R}(\boldsymbol{\mu}_i(\boldsymbol{\beta}, d) \Rightarrow \mathbf{B}(\boldsymbol{\beta}, d))\}\mathrm{L}_{\boldsymbol{\mu}_i(\boldsymbol{\beta},d)}(\mathbf{m}_i(d))$, where $\boldsymbol{\mu}_i(\boldsymbol{\beta}, d)$ is the conditional mean of $\mathbf{m}_i(d)$ given $\mathbf{x}_i$. Model (12) leads to an intrinsic regression model for $M(1)^K$ given by

$$E\{\mathcal{E}_i(\boldsymbol{\beta}, d) \mid \mathbf{x}_i\} = \mathbf{0} \tag{13}$$

for all $d \in \mathcal{D}$ and $i = 1, \ldots, n$. As a comparison, consider a multivariate regression model $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ and $E(\boldsymbol{\epsilon}_i \mid \mathbf{x}_i) = E(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} \mid \mathbf{x}_i) = \mathbf{0}$, where $\mathbf{Y}_i$ is a $p_y \times 1$ vector and $\mathbf{X}_i$ is a $p_y \times p$ design matrix depending on $\mathbf{x}_i$. It is clear that $\mathcal{E}_i(\boldsymbol{\beta}, d)$ is closely related to $\boldsymbol{\epsilon}_i = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}$ in the multivariate regression model and thus the intrinsic regression model (13) for $M(1)^K$ can be regarded as a generalization of a standard multivariate regression.

The key advantage of translating tangent vectors on different tangent spaces to the same tangent space is that we can directly apply most multivariate analysis techniques in Euclidean space to the analysis of $\mathcal{E}_i(\boldsymbol{\beta})$ (Anderson, 2003). By using parallel transport to obtain $\mathcal{E}_i(\boldsymbol{\beta})$, we can explicitly account for correlation structure among $\mathcal{E}_i(\boldsymbol{\beta})$ and then construct a set of

estimation equations to calculate a more efficient parameter estimate. Please refer to the next section for details.

2.3 Two-stage estimation procedure

We propose a two-stage estimation procedure for computing parameter estimates for the semi-parametric medial representation regression model (12) as follows.

Stage 1 is to calculate an intrinsic least squares estimate of the parameter $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_I$, by minimizing the square of the geodesic distance,

$$\hat{\boldsymbol{\beta}}_I = \text{argmin}_{\boldsymbol{\beta}} D_n(\boldsymbol{\beta}) = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} D_{n,i}(\boldsymbol{\beta}) = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \text{dist}\{\mathbf{m}_i, \boldsymbol{\mu}_i(\boldsymbol{\beta})\}^2, \qquad (14)$$

where $D_{n,i}(\boldsymbol{\beta}) = \text{dist}\{\mathbf{m}_i, \boldsymbol{\mu}_i(\boldsymbol{\beta})\}^2$ and $\text{dist}\{\mathbf{m}_i, \boldsymbol{\mu}_i(\boldsymbol{\beta})\}$ is the shortest distance between $\mathbf{m}_i$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ on $M(1)$. Since $D_n(\boldsymbol{\beta})$ can be written as the sum of four terms: $D_n^{(1)}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\{\mathbf{O}_i - \boldsymbol{\mu}_{oi}(\boldsymbol{\beta})\}^T\{\mathbf{O}_i - \boldsymbol{\mu}_{oi}(\boldsymbol{\beta})\}$, $D_n^{(2)}(\boldsymbol{\beta}) = \sum_{i=1}^{n}[\log(r_i) - \log\{\mu_{ri}(\boldsymbol{\beta})\}]^2$, $D_n^{(3)}(\boldsymbol{\beta}) = \sum_{i=1}^{n}[\arccos\{\mathbf{s}_{0i}^T\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})\}]^2$ and $D_n^{(4)}(\boldsymbol{\beta}) = \sum_{i=1}^{n}[\arccos\{\mathbf{s}_{1i}^T\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})\}]^2$, we can minimize $D_n^{(k)}(\boldsymbol{\beta})$ for $k = 1, 2, 3, 4$ independently when they do not share any common parameters.

Computationally, we develop an annealing evolutionary stochastic approximation Monte Carlo algorithm (Liang, 2011) for obtaining $\hat{\boldsymbol{\beta}}_I$, whose details can be found in the supplementary report. Moreover, according to our experience, the traditional optimization methods including the quasi-Newton method do not perform well for optimizing $D_n(\boldsymbol{\beta})$ and strongly depend on the starting value of $\boldsymbol{\beta}$. When $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ takes a relatively complicated form, $D_n(\boldsymbol{\beta})$ is generally not concave and can have multiple local modes. For instance, since $\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})$ is a nonlinear function of $\boldsymbol{\beta}$ and $D_n^{(4)}(\boldsymbol{\beta})$ may not be a concave function of $\boldsymbol{\beta}$ over $\mathcal{B}$, our prior experiences have shown that the quasi-Newton method for optimizing $D_n^{(4)}(\boldsymbol{\beta})$ can easily converge to local minima.

The estimate $\hat{\boldsymbol{\beta}}_I$ is closely associated with the intrinsic mean (Bhattacharya and Patrangenaru, 2005) and does not involve the concept of parallel transport. If we replace $|\arccos(s)|^2$ by $1 - s$ in $D_n^{(3)}(\boldsymbol{\beta})$ and $D_n^{(4)}(\boldsymbol{\beta})$, then our fitting procedure in Stage 1 is effectively a maximum

likelihood estimation for a model with the Fisher-distributed errors on the sphere and thus $\hat{\boldsymbol{\beta}}_I$ is an extrinsic estimate. It will be shown in Theorem 1 below that $\hat{\boldsymbol{\beta}}_I$ is a consistent estimate, but $\hat{\boldsymbol{\beta}}_I$ is not efficient, since it does not account for the correlation among the different components of medial representations.

Stage 2 is to calculate a more efficient estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_E$, which is a solution of

$$\sum_{i=1}^{n} \hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\mathcal{E}_i(\boldsymbol{\beta}) = \mathbf{0}, \tag{15}$$

where $\hat{\mathbf{h}}_E(\mathbf{x}_i) = \partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_I)\{\mathbf{R}(\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_I) \Rightarrow \mathbf{B}(\hat{\boldsymbol{\beta}}_I))\}^{-1} = \partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_I)\{\mathbf{R}(\mathbf{B}(\hat{\boldsymbol{\beta}}_I) \Rightarrow \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_I))\}$, $\mathbf{V}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{E}_i(\boldsymbol{\beta})\mathcal{E}_i(\boldsymbol{\beta})^T/n$, and $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\beta}}_I)$.

The equation (15) in Stage 2 is invariant to the rotation matrix $\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)$, where $\mathbf{P}_0 = (0, 0, 0, 1, 0, 0, 1, 0, 0, 1)^T$ representing the center at the origin $(0, 0, 0)^T$, the unit radius $r = 1$, and the two spoke directions pointing towards the north pole $(0, 0, 1)^T$. Specifically, we can use the rotation matrix $\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)$ to rotate $\mathcal{E}_i(\boldsymbol{\beta})$ to $\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}\mathcal{E}_i(\boldsymbol{\beta})$ for all i. Correspondingly, $\hat{\mathbf{h}}_E(\mathbf{x}_i)$ and $\mathbf{V}^{-1}$ are, respectively, changed to $\hat{\mathbf{h}}_E(\mathbf{x}_i)\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}^T$ and $\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}\mathbf{V}^{-1}\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}^T$. Thus, after applying the rotation $\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)$, we can show that $\hat{\mathbf{h}}_E(\mathbf{x}_i)\mathbf{V}^{-1}\mathcal{E}_i(\boldsymbol{\beta})$ equals

$$\hat{\mathbf{h}}_E(\mathbf{x}_i)\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}^T\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}\mathbf{V}^{-1}\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}^T\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)\}\mathcal{E}_i(\boldsymbol{\beta}),$$

which is independent of $\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \mathbf{P}_0)$.

Model (12) is a conditional mean model (Chamberlain, 1987; Newey, 1993). The conditional mean model implies that $E\{\mathbf{h}(\mathbf{x}_i)\mathcal{E}_i(\boldsymbol{\beta})\} = E[\mathbf{h}(\mathbf{x}_i)E\{\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\}] = \mathbf{0}$ for any vector function $\mathbf{h}(\cdot)$, which may depend on $\boldsymbol{\beta}$. After some algebraic calculations, it can be shown that calculating $\hat{\boldsymbol{\beta}}_I$ is equivalent to solving $\partial_{\boldsymbol{\beta}} D_n(\boldsymbol{\beta}) = -2\sum_{i=1}^{n} \partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\boldsymbol{\beta})\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \boldsymbol{\mu}_i(\boldsymbol{\beta}))\mathcal{E}_i(\boldsymbol{\beta}) = \mathbf{0}$, that is, $\mathbf{h}_I(x_i) = \partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\boldsymbol{\beta})\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \boldsymbol{\mu}_i(\boldsymbol{\beta}))$. However, it has been shown (Chamberlain, 1987; Newey, 1993) that the optimal function has the form $\mathbf{h}_{opt}(\mathbf{x}_i, \boldsymbol{\beta}) = E\{\partial_{\boldsymbol{\beta}}\mathcal{E}_i(\boldsymbol{\beta}) \mid x_i\}\text{var}\{\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\}^{-1}$, which achieves the semiparametric efficiency bound for $\boldsymbol{\beta}$. Therefore, $\mathbf{h}_I(\mathbf{x}_i)$ is not an optimal function and thus the intrinsic least squares estimate in Stage 1 is not an efficient estimator.

Since $E\{\partial_{\boldsymbol{\beta}}\mathcal{E}_i(\boldsymbol{\beta}) \mid x_i\}$ and $\text{var}\{\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\}$ for each $\boldsymbol{\beta}$ do not have a simple form, we must estimate them nonparametrically, which leads to a nonparametric estimate of $\mathbf{h}_{opt}(\mathbf{x}, \boldsymbol{\beta})$, denoted by $\hat{\mathbf{h}}_{opt}(\mathbf{x}, \boldsymbol{\beta})$. Although we may solve the estimating equations $\mathbf{F}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \hat{\mathbf{h}}_{opt}(x_i, \boldsymbol{\beta})\mathcal{E}_i(\boldsymbol{\beta}) = \mathbf{0}$ to calculate the efficient estimator of $\boldsymbol{\beta}$, it can be computationally challenging to solve $\mathrm{F}_n(\boldsymbol{\beta})$ since nonparametrically, estimating the $8 \times p$ matrix $E\{\partial_{\boldsymbol{\beta}}\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\}$ and the 8×8 inverse matrix of $\text{var}\{\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\}$ can be very unstable for a relatively small sample size. Thus, we replace $\text{var}\{\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\}$ by $\text{var}\{\mathcal{E}_i(\boldsymbol{\beta})\}$ and approximate $E\{\partial_{\boldsymbol{\beta}}\mathcal{E}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i\}$ by $\partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\boldsymbol{\beta})\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \boldsymbol{\mu}_i(\boldsymbol{\beta}))$. Moreover, in order to avoid calculating $\partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\boldsymbol{\beta})\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \boldsymbol{\mu}_i(\boldsymbol{\beta}))$ and $\text{var}\{\mathcal{E}_i(\boldsymbol{\beta})\}$ during each numerical iteration, we calculate them at $\hat{\boldsymbol{\beta}}_I$ and then construct the objective function $\sum_{i=1}^n \hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\mathcal{E}_i(\boldsymbol{\beta}) = \mathbf{0}$ for calculating $\hat{\boldsymbol{\beta}}_E$. The two-stage estimation procedure leads to substantial computational efficiency, since solving the complex estimating equations (15) is relatively easy starting from $\hat{\boldsymbol{\beta}}_I$. An alternative way is to directly minimize $\{\sum_{i=1}^n \partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\boldsymbol{\beta})\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}) \Rightarrow \boldsymbol{\mu}_i(\boldsymbol{\beta}))\mathbf{V}(\boldsymbol{\beta})^{-1}\mathcal{E}_i(\boldsymbol{\beta})\}^2$, which is much more complex than $D_n(\boldsymbol{\beta})$ and thus is computationally difficult.

As a comparison between $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_I$, we consider a multivariate nonlinear regression model $\mathbf{Y}_i = \mathbf{F}(\mathbf{x}_i, \boldsymbol{\beta}) + \boldsymbol{\epsilon}_i$ with $E(\boldsymbol{\epsilon}_i \mid \mathbf{x}_i) = E\{\mathbf{Y}_i - \mathbf{F}(\mathbf{x}_i, \boldsymbol{\beta}) \mid \mathbf{x}_i\} = \mathbf{0}$ and $\text{var}(\boldsymbol{\epsilon}_i \mid \mathbf{x}_i) = \boldsymbol{\Sigma}$, where $\mathbf{F}(\mathbf{x}_i, \boldsymbol{\beta})$ is a vector of nonlinear functions of $\mathbf{x}_i$ and $\boldsymbol{\beta}$. In this case, $\mathcal{E}_i(\boldsymbol{\beta}) = \boldsymbol{\epsilon}_i = \mathbf{Y}_i - \mathbf{F}(\mathbf{x}_i, \boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_I = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n \{\mathbf{Y}_i - \mathbf{F}(\mathbf{x}_i, \boldsymbol{\beta})\}^T\{\mathbf{Y}_i - \mathbf{F}(\mathbf{x}_i, \boldsymbol{\beta})\}$, and $\hat{\mathbf{h}}_E(\mathbf{x}_i) = \partial_{\boldsymbol{\beta}}\mathbf{F}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_I)$. Then, $\boldsymbol{\Sigma}$ can be estimated by using $\hat{\mathbf{V}} = \sum_{i=1}^n \{\mathbf{Y}_i - \mathbf{F}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_I)\}\{\mathbf{Y}_i - \mathbf{F}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_I)\}^T/n$. Equation (15) reduces to $\sum_{i=1}^n \hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\{\mathbf{Y}_i - \mathbf{F}(\mathbf{x}_i, \boldsymbol{\beta})\} = \mathbf{0}$, whose solution is just $\hat{\boldsymbol{\beta}}_E$. Under mild conditions, it can be shown that compared with $\hat{\boldsymbol{\beta}}_I$, $\hat{\boldsymbol{\beta}}_E$ is a more efficient estimator of $\boldsymbol{\beta}$ and its asymptotic covariance is given by $\{\sum_{i=1}^n \hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\hat{\mathbf{h}}_E(\mathbf{x}_i)^T\}^{-1}$. In the context of highly concentrated spoke data, our intrinsic regression model reduces to the multivariate nonlinear regression model and similar to the multivariate nonlinear regression model, the two-stage approach can increase statistical efficiency in estimating $\boldsymbol{\beta}$.

2.4 Asymptotic properties

We establish consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_I$ and $\hat{\boldsymbol{\beta}}_E$. The following assumptions are needed to facilitate the technical details, although they are not the weakest possible conditions.

Assumption A1. The data $\{\mathbf{z}_i = (\mathbf{x}_i, \mathbf{m}_i) : i = 1, \cdots, n\}$ form an independent and identical sequence.

Assumption A2. $\boldsymbol{\beta}_*$ is an interior point of the compact set $\mathcal{B} \subset R^p$ and is the unique solution for the model, $E\{\mathbf{h}_E(\mathbf{x})\mathcal{E}(\boldsymbol{\beta})\} = \mathbf{0}$, where $\mathbf{h}_E(x) = \partial_{\boldsymbol{\beta}}\boldsymbol{\mu}_i(\boldsymbol{\beta}_*)\{\mathbf{R}(\mathbf{B}(\boldsymbol{\beta}_*) \Rightarrow \boldsymbol{\mu}_i(\boldsymbol{\beta}_*))\}\mathbf{V}(\boldsymbol{\beta}_*)^{-1}$. Moreover, $\boldsymbol{\beta}_*$ is an isolated point of the set of all minimizers of the map $D(\boldsymbol{\beta}) = E[\text{dist}\{\mathbf{m}, \boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta})\}^2]$ on $\mathcal{B}$, denoted by $I_{\mathcal{B}}$.

Assumption A3. In an open neighborhood of $\boldsymbol{\beta}_*$, $\boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta})$ has a second-order continuous derivative with respect to $\boldsymbol{\beta}$ and $||\mathrm{L}_{\mu(\boldsymbol{\beta})}(\mathbf{m})||$, $||\partial_\mu \mathrm{L}_{\mu(\boldsymbol{\beta})}(m)||$, $||\partial_{\boldsymbol{\beta}}\mu(\mathbf{x}, \boldsymbol{\beta})||$ and $||\partial^2_{\boldsymbol{\beta}}\mu(\mathbf{x}, \boldsymbol{\beta})||$ are bounded by some integrable function $G(\mathbf{z})$ with $\mathrm{E}\{G(\mathbf{z})^2\} < \infty$.

Assumption A4. In an open neighborhood of $\boldsymbol{\beta}_*$, the rank of $E\left\{\partial^2_{\boldsymbol{\beta}} D_{n,i}(\boldsymbol{\beta})\right\}$ is p and $E[\left\{\partial_{\boldsymbol{\beta}} D_{n,i}(\boldsymbol{\beta})\right\}^{\otimes 2}]$ is positive definite, where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for a given vector $\mathbf{a}$.

Assumption A1 is needed just for notational simplicity and can be easily modified to accommodate independent and non-identically distributed scenarios. Assumption A2 is an identifiability condition. Assumptions A3 and A4 are standard conditions for ensuring the first order asymptotic properties including consistency and asymptotic normality of M-estimators when the sample size is large (van der Vaart and Wellner, 1996). We obtain the following theorems, whose detailed proofs can be found in the Appendix.

Theorem 1. *(a) If assumptions A1, A2, and A3 are true, then $\hat{\boldsymbol{\beta}}_I$ and $\hat{\boldsymbol{\beta}}_E$ converge to $\boldsymbol{\beta}_*$ in probability as $n \to \infty$, where $\boldsymbol{\beta}_*$ is the solution of (12).*

(b) Under assumptions A1-A4, we have

$$[E\sum_{i=1}^{n}\{\partial_{\boldsymbol{\beta}} D_{n,i}(\hat{\boldsymbol{\beta}}_I)^{\otimes 2}\}]^{-1/2} E\{-\partial^2_{\boldsymbol{\beta}} D_n(\hat{\boldsymbol{\beta}}_I)\}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_*) \to N(\mathbf{0}, \mathbf{I}_p) \tag{16}$$

as $n \to \infty$, where I_p is a $p \times p$ identity matrix and $\to$ denotes convergence in distribution.

(c) Under assumptions A1-A4, we have

$$[\sum_{i=1}^{n}\{\hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\mathcal{E}_i(\hat{\boldsymbol{\beta}}_E)\}^{\otimes 2}]^{-1/2}\{\sum_{i=1}^{n}\hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\partial_{\boldsymbol{\beta}}\mathcal{E}_i(\hat{\boldsymbol{\beta}}_E)^T\}(\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_*) \to N(\mathbf{0}, \mathbf{I}_p) \tag{17}$$

as $n \to \infty$.

Theorem 1 has several important applications. Theorem 1 (a) establishes the consistency of $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_I$. According to Theorems 1 (b) and (c), we can consistently estimate the covariance matrices of $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_I$. For instance, the covariance matrix of $\hat{\boldsymbol{\beta}}_E$, denoted by $\hat{\boldsymbol{\Sigma}}_E$, can be approximated by

$$\{\sum_{i=1}^{n}\hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\partial_{\boldsymbol{\beta}}\mathcal{E}_i(\hat{\boldsymbol{\beta}}_E)^T\}^{-1}[\sum_{i=1}^{n}\{\hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\mathcal{E}_i(\hat{\boldsymbol{\beta}}_E)\}^{\otimes 2}]\{\sum_{i=1}^{n}\hat{\mathbf{h}}_E(\mathbf{x}_i)\hat{\mathbf{V}}^{-1}\partial_{\boldsymbol{\beta}}\mathcal{E}_i(\hat{\boldsymbol{\beta}}_E)^T\}^{-T}. \tag{18}$$

Moreover, we can use Theorem 1 (c) to construct confidence cones of $\hat{\boldsymbol{\beta}}_E$ and its functions. Since Theorem 1 only establishes the asymptotic properties of $\hat{\boldsymbol{\beta}}_E$ when the sample size is large, these properties may be inadequate to characterize the finite sample behavior of $\hat{\boldsymbol{\beta}}_E$ for relatively small samples. In the case of small samples, we may have to resort to higher order approximations, such as saddlepoint approximations and bootstrap methods (Butler, 2007; Davison and Hinkley, 1997).

Our choices of which hypotheses to test are motivated by scientific questions, which involve a comparison of medial representation components across diagnostic groups. These questions usually can be formulated as testing linear hypotheses of $\boldsymbol{\beta}$ as follows:

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b}_0 \quad \text{vs.} \quad H_1 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{b}_0, \tag{19}$$

where $\mathbf{A}$ is an $r \times p$ matrix of full row rank and $\mathbf{b}_0$ is an $r \times 1$ specified vector. We test the null hypothesis $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b}_0$ using a Wald test statistic W_n defined by

$$W_n = (\mathbf{A}\hat{\boldsymbol{\beta}}_E - \mathbf{b}_0)^T(\mathbf{A}\hat{\boldsymbol{\Sigma}}_E\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}}_E - \mathbf{b}_0). \tag{20}$$

We are led to the following theorem.

Theorem 2. *If the assumptions A1-A4 are true, then the statistic W_n is asymptotically distributed as $\chi^2(r)$, a chi-square distribution with r degrees of freedom, under the null hypothesis H_0.*

An asymptotically valid test can be obtained by comparing sample values of the test statistic with the critical value of a $\chi^2(r)$ distribution at a pre-specified significance level α. However, for a small sample size n, we observed relatively low precision of the chi-square approximation. Instead, we calibrate W_n with a critical value of $F^{1-\alpha}_{r,n-r} r(n-1)/(n-r)$, which leads to a slightly higher precision of the F approximation, where $F^{1-\alpha}_{r,n-r}$ is the upper α-percentile of the $F_{r,n-r}$ distribution. That is, we reject H_0 if $W_n \geq F^{1-\alpha}_{r,n-r} r(n-1)/(n-r)$, and do not reject H_0 otherwise. The reason that the F approximation outperforms the chi-square approximation is due to the fact that the F approximation explicitly accounts for sample uncertainty in estimating the covariance matrix of $A\hat{\boldsymbol{\beta}}_E$.

3 Simulation studies and real data

3.1 Double directional data with covariates

We generated double directional responses as follows:

$$\mathbf{R}_{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}),(0,0,-1)^T} \mathrm{L}_{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})}(\mathbf{s}_{0i}) = \mathcal{E}_{0i}, \quad \mathbf{R}_{\boldsymbol{\mu}_{1i}(\boldsymbol{\beta}),(0,0,-1)^T} \mathrm{L}_{\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})}(\mathbf{s}_{1i}) = \mathcal{E}_{1i},$$

where $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ and $\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})$ were set according to (6), in which $\mathbf{x}_{i,d}$'s were fixed at 1 and $\mathbf{x}_{i,c}$'s were independently simulated from a $N(0,1)$ distribution. It is assumed that both $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ and $\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})$ were, respectively, centered around $\mathbf{g}_5(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{5d}) = (u_0, v_0, w_0)^T$ and $\mathbf{g}_6(\mathbf{x}_{i,d}, \boldsymbol{\beta}_{6d}) = (u_1, v_1, w_1)^T$ according to (1) such that

$$\frac{u_0}{1-w_0} = \beta_{5d,1} = 1.2, \ \frac{v_0}{1-w_0} = \beta_{5d,2} = 1.2, \ \frac{u_1}{1-w_1} = \beta_{6d,1} = 0.8, \text{ and } \frac{v_1}{1-w_1} = \beta_{6d,2} = 0.8.$$

In addition, we imposed two constraints as follows:

$$\boldsymbol{\beta}_{5c} = (\beta_{5c,1}, \beta_{5c,2})^T = \boldsymbol{\beta}_{6c} = (\beta_{6c,1}, \beta_{6c,2})^T = (1,1)^T.$$

We generated the errors $\mathcal{E}_{0i}$ and $\mathcal{E}_{1i}$ in $T_{(0,0,-1)}(S^2)$ from a 4-dimensional normal distribution, $N(0, 0.5\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being specified as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_{01} \\ \boldsymbol{\Sigma}_{01} & \boldsymbol{\Sigma}_1 \end{pmatrix}, \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_{01} = \rho_2 \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}.$$

Subsequently, we rotated $\mathcal{E}_{0i}$ onto the tangent space $T_{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})}(S^2)$ and $\mathcal{E}_{1i}$ onto the tangent space $T_{\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})}(S^2)$, and then we used the Exp map defined in the supplementary report to obtain the responses $\mathbf{s}_{0i}$ and $\mathbf{s}_{1i}$. We set $n = 40$, 80, and 120, $\rho_1 = \rho_2 = 0.5$, and then we simulated 2000 datasets for each case to compare the biases and the root-mean-square error of the two estimates: $\hat{\boldsymbol{\beta}}_I$ and $\hat{\boldsymbol{\beta}}_E$. As seen in Table 1, $\hat{\boldsymbol{\beta}}_E$ has smaller root-mean-square error than $\hat{\boldsymbol{\beta}}_I$ for every component of $\boldsymbol{\beta}$, but some components of $\hat{\boldsymbol{\beta}}_E$ can be more biased.

We also calculated the mean of the estimated standard error estimates and the relative efficiencies for all the components in $\hat{\boldsymbol{\beta}}_E$ and evaluated the finite sample performance of the Wald statistic W_n for hypothesis testing. The results are quite similar to those from the single directional case in the supplementary file, so we did not present them here to preserve space.

3.2 Schizophrenia study of the hippocampus

We consider a neuroimaging dataset about the medial representation shape of the hippocampus structure in the left and right brain hemisphere in schizophrenia patients and healthy controls, collected at 14 academic medical centers in North America and western Europe. The hippocampus, a gray matter structure in the limbic system, is involved in processes of motivation and emotions, and plays a central role in the formation of memory.

In this study, 238 first-episode schizophrenia patients (53 female, 185 male; mean/standard deviation age, female 25.1/5.69 years; male 23.6/4.55 years) were enrolled who met the following criteria: age 16 to 40 years; onset of psychiatric symptoms before age 35; diagnosis of schizophrenia, schizophreniform, or schizoaffective disorder according to DSM-IV criteria; and

Table 1: Bias ($\times 10^{-3}$) and MS ($\times 10^{-2}$) of $\hat{\boldsymbol{\beta}}_I$ and $\hat{\boldsymbol{\beta}}_E$ for double directional case. Bias denotes the bias of the mean of the estimates; MS denotes the root-mean-square error. For each parameter, the first row is for $\hat{\boldsymbol{\beta}}_I$ and the second is for $\hat{\boldsymbol{\beta}}_E$. Moreover, the constraints $\beta_{5c,1} = \beta_{6c,1}$ and $\beta_{5c,2} = \beta_{6c,2}$ are imposed.

	$n = 40$		$n = 80$		$n = 120$	
	Bias	MS	Bias	MS	Bias	MS
$\boldsymbol{\beta}_{5d,1} = 1.2$	3.15	13.26	4.35	10.04	4.22	7.75
	3.40	13.10	4.36	9.82	3.98	7.60
$\boldsymbol{\beta}_{5c,1} = \boldsymbol{\beta}_{6c,1} = 1$	9.29	19.19	1.74	12.76	7.43	10.31
	8.93	18.02	0.89	12.09	7.27	9.81
$\boldsymbol{\beta}_{5d,2} = 1.2$	9.44	13.69	2.05	10.19	0.86	7.80
	9.81	13.29	0.88	9.59	0.43	7.69
$\boldsymbol{\beta}_{5c,2} = \boldsymbol{\beta}_{6c,2} = 1$	6.90	18.55	5.00	13.08	0.64	10.53
	6.74	17.50	5.67	12.44	0.62	9.99
$\boldsymbol{\beta}_{6d,1} = 0.8$	5.18	16.85	3.23	9.74	2.49	7.93
	5.69	12.91	3.10	9.65	2.69	7.76
$\boldsymbol{\beta}_{6d,2} = 0.8$	2.34	14.84	1.31	9.78	0.86	8.47
	1.32	13.06	0.98	9.71	0.91	8.07

various treatment and substance dependence conditions. 56 healthy control subjects (18 female, 38 male; mean/standard deviation age, female 24.8/3.30 years; male 25.3/4.21 years) were also enrolled. Neurocognitive and magnetic resonance imaging (MRI) assessments were performed at the first visit time.

The brain MRI data were first aligned to the Montreal Neurological Institute (MNI) space. Hippocampi were segmented in the MNI space and then their medial representations were reconstructed from those binary segmentations (Styner et al., 2004). Subsequently, these hippocampus medial representations were realigned by using a rigid body variation of the standard Procrustes method. The resulting alignment leads to a shape representation that is invariant to translation and rotation, but not to scale. Scaling information is retained for studying changes in overall size or volume.

The aim of our study was to investigate the difference of medial representation shape between schizophrenia patients and healthy controls while controlling for other factors, such as gender and age. The response of interest was the hippocampus medial representation shape at the 24 medial atoms of the left and right brain hemisphere (Figure 1). Covariates of interest were Whole Brain Volume (WBV), race including Caucasian, African American and others, age in years, gender, and diagnostic status including patient and control.

The covariate vector is $\mathbf{x}_i = (1, \text{gender}_i, \text{age}_i, \text{diag}_i, \text{race1}_i, \text{race2}_i, \text{WBV}_i)^T$, where diag is the dummy variable for patients versus healthy controls, and race1 and race2 are, respectively, dummy variables for Caucasians and African Americans versus other races. For the location component on the medial representation, we set $\mu_O(\mathbf{x}, \boldsymbol{\beta}) = (\mathbf{x}^T\boldsymbol{\beta}_1, \mathbf{x}^T\boldsymbol{\beta}_2, \mathbf{x}^T\boldsymbol{\beta}_3)^T$, where $\boldsymbol{\beta}_k$ $(k = 1, 2, 3)$ are 7×1 coefficient vectors. For the radius component on the medial representation, we set $\mu_r(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}^T\boldsymbol{\beta}_4)$, where $\boldsymbol{\beta}_4$ is a 7×1 coefficient vector. For the directional components on the medial representation, we used $\boldsymbol{\mu}_0(\mathbf{x}_i, \boldsymbol{\beta})$ as defined in (6), in which $\mathbf{x}_{i,d} = (\text{gender}_i, \text{diag}_i, \text{race1}_i, \text{race2}_i)^T$, $\mathbf{x}_{i,c} = (\text{age}_i, \text{WBV}_i)^T$, $\boldsymbol{\beta}_5 = (\boldsymbol{\beta}_{5d}^T, \boldsymbol{\beta}_{5c}^T)^T$ for s_0 and $\boldsymbol{\beta}_6 = (\boldsymbol{\beta}_{6d}^T, \boldsymbol{\beta}_{6c}^T)^T$ for $\mathbf{s}_1$. Therefore, we have the coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T, \boldsymbol{\beta}_4^T, \boldsymbol{\beta}_5^T, \boldsymbol{\beta}_6^T)^T$.

Then we used the two-stage estimation procedure to obtain estimates of $\boldsymbol{\beta}$ and conducted hypothesis testing using Wald statistics. Since the primary goal of the study is to investigate the difference of medial representation shape between schizophrenia patients and healthy controls, we paid special attention to the terms in $\boldsymbol{\beta}$ associated with diagnostic status.

First, we examined the overall diagnostic status effect on the whole medial representation structure. The p-values of the diagnostic status effects across the atoms of both the left and right reference hippocampi are shown in the first row (a) and (b) of Figure 3. The false discovery rate approach (Benjamini and Hochberg, 1995) was used to correct for multiple comparisons, and the corresponding adjusted p-values are shown in the first row (c) and (d) of Figure 3. There was a large significant area in the left hippocampus and also some in the right hippocampus. The significance area remains almost the same after correcting for multiple comparisons, but with an attenuated significance level.

We also examined each component on the medial representation separately. For the radius component of the medial representation, we presented the p-values of the diagnostic status effects across the atoms in the second row (a) and (b) of Figure 3 and the adjusted p-values in the second row (c) and (d). Before correcting for multiple comparisons, we observed a significant diagnostic status difference in the medial representation thickness at the central atoms near the posterior side in the left hippocampus and in some areas in the right hippocampus, whereas we did not observe much of a significant diagnostic status effect after correcting for multiple comparisons.

For the location component of the medial representation, we showed the p-values of the diagnostic status effects in the third row (a) and (b) of Figure 3 and the corresponding adjusted p-values in the third row (c) and (d). We observed significant diagnostic status differences mainly located around the anterior and lateral side of the left hippocampus though with clearly reduced significance after correcting for multiple comparisons. Similar lateral results have also been observed by Narr et al. (2004).

Similarly, for the two spoke directions on the medial representation, the p-values of the di-

agnostic status effects are shown in the last row (a) and (b) of Figure 3 and the corresponding adjusted p-values are shown in the last row (c) and (d). Before correcting for multiple comparisons, there was some significant area around the anterior, posterior, and the medial side of the left hippocampus, but not much in the right hippocampus. There was still some significance for the diagnostic status effect around the same areas in the left hippocampus after correcting for multiple comparisons, but nothing in the right hippocampus. The posterior orientation effect of hippocampal differences in schizophrenia has also been shown by Styner et al. (2004) and basically constitutes a local bending change in that region. The anterior effect is novel and located at the intersection of the hippocampal Cornu Ammonis 1 and Cornu Ammonis 2 regions.

We also examined the overall age effect on the whole medial representation structure. The color-coded p-values of the age effect across the atoms of both the left and right reference hippocampi are shown in the first row (a) and (b) of Figure 4. The false discovery rate approach was used to correct for multiple comparisons, and the corresponding adjusted p-values are shown in the first row (c) and (d) of Figure 4. There was a large significant area in the right hippocampus and also some in the left hippocampus. The significance area remains almost the same after correcting for multiple comparisons, but with an attenuated significance level.

Additionally, we looked at each component on the medial representation separately. For the radius component of the medial representation, the color-coded p-values of the age effect across the atoms are shown in the second row (a) and (b) of Figure 4 and the adjusted p-values are shown in the second row (c) and (d). Before correcting for multiple comparisons, there was a small age effect in the medial representation thickness at the central atoms near the posterior side in the left hippocampus and in some areas in the right hippocampus. However, there was not much of a significant diagnostic status effect after correcting for multiple comparisons.

For the location component of the medial representation, the color-coded p-values of the age effect are shown in the third row (a) and (b) of Figure 4 and the corresponding adjusted p-values are shown in the third row (c) and (d). Significant age effects were mainly located around the

anterior and lateral side of the left hippocampus though with clearly reduced significance after correcting for multiple comparisons.

For the two spoke directions on the medial representation, we showed the color-coded p-values of the age effect in the last row (a) and (b) of Figure 4 and the corresponding adjusted p-values are in the last row (c) and (d). Even after correcting for multiple comparisons, we observed significant areas around the anterior, posterior, and the medial side of the right hippocampus and some areas in the left hippocampus.

Finally, following suggestions from a reviewer, we examined the overall diagnostic status effect without accounting for other factors. The p-values of the diagnostic status effects are shown in Figure 5. Inspecting Figure 5 reveals a small significant area in the left and right hippocampi before and after correcting for multiple comparisons. Comparing with Figure 3, we feel that such attenuation in Figure 5 may be caused by omitting other factors such as age that are believed to be associated with the variability of the medial representation of subcortical structures.

4 Discussion

We have proposed a semiparametric model for describing the association between the medial representation of subcortical structures and covariates of interest, such as diagnostic status, age and gender. We have developed a two-stage estimation procedure to calculate the parameter estimates and used Wald statistics to test linear hypotheses of unknown parameters. We have used extensive simulation studies and a real dataset to evaluate the accuracy of our parameter estimates and the finite sample performance of the Wald statistics.

Many issues still merit further research. The two-stage estimation procedure can be easily modified to simultaneously estimate all parameters across all atoms and imposing some structures (e.g., spatial smoothness) on the matrix of regression parameters across all atoms while accounting for the correlations between different components of different atoms. This general-

ization requires a good estimate of the covariance matrix of $\mathcal{E}_i(\boldsymbol{\beta})$ across all atoms. We may consider a shrinkage estimator of the covariance matrix of all $\mathcal{E}_i(\boldsymbol{\beta})$ as a linear combination of the identity matrix and the sample covariance matrix $\mathbf{V}(\boldsymbol{\beta})$ (Ledoit and Wolf, 2004). Moreover, for the matrix of regression parameters across all atoms, we may consider its sparse low-rank matrix factorization to identify the underlying latent structure among all atoms (Witten, Tibshirani, and Hastie, 2009; Dryden and Mardia, 1998; Fletcher et al., 2004), which will be a topic of our future research. It is interesting to develop Bayesian models for the joint analysis of medial representation data of subcortical structures (Angers and Kim, 2005; Healy and Kim, 1996).

References

Andrews, D. W. K. (1992), "Generic uniform convergence," *Econometric Theory*, 8, 241-257.

Andrews, D. W. K. (1994), "Empirical Process Methods in Econometrics," *Handbook of Econometrics*, Volume IV. Edited by Engle, R. F. and McFadden, D. L., 2248-2292.

Andrews, D. W. K. (1999), "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67, 543-564.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.), Wiley Series in Probability and Statistics.

Angers, J. F. and Kim, P. T. (2005), "Multivariate Bayesian Function Estimation," *Ann. Statist.*, 33, 2967-2999.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B 57, 289-300.

Bhattacharya, R. N. and Patrangenaru, V. (2005), "Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds II," *Ann. Statist*, 33, 1225-1259.

Butler, R. W. (2007). *Saddlepoint Approximations with Applications*. New York, Cambridge University Press.

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *J. Economet.*, 34, 305-334.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. New York, Cambridge University Press.

Downs, T. D. (2003), "Spherical Regression," *Biometrika*, 90, 655-668.

Dryden, I. L and Mardia, K. V. . (1998). *Statistical Shape Analysis*. Wiley, Chichester.

Fisher, N. I. and Lee, A. J. (1992), "Regression Models for an Angular Response," *Biometrics*, 48, 665-677.

Fletcher P. T., Lu C., Pizer S. M. and Joshi S. (2004), "Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape," *Medical Imaging*, 23, 995-1005.

Gould, A. L. (1969), "A Regression Technique for Angular Variates," *Biometrics*, 25, 683-700.

Healy, D. M. and Kim, P. T. (1996), "An Empirical Bayes Approach to Directional Data and Efficient Computation on the Sphere," *Ann. Statist.*, 24 232–254.

Jennrich R. (1969), "Asymptotic Properties of Nonlinear Least Squares Estimators," *Ann. of Math. Statist.*, 40, 633-643

Johnson, R. A. and Wehrly, T. E. (1978), "Some Angular-linear Distributions and Related Regression Models," *J. Am. Statist. Assoc.*, 73, 602-606.

Jupp, P. E. and Mardia, K. V. (1989), "A Unified View of the Theory of Directional Statistics, 1975-1988," *International Statistical Review*, 57, 261-294.

Le, H. (2001), "Locating Frechet means with an application to shape spaces," *Adv. Appl. Prob.*, 33, 324-338.

Ledoit, O. and Wolf, M. (2004), "A Well-conditioned Estimator for Large-dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365-411.

Liang, F. (2011), "Annealing Evolutionary Stochastic Approximation Monte Carlo for Global Optimization," *Statistics and Computing*, 21, 375-393.

Mardia, K. V. (1975), "Statistics of Directional Data (with Discussion)," *J. R. Statist. Soc. B*, 37, 349-393.

Mardia, K. V. and Jupp, P. E. (1983). *Directional Statistics*, Academic Press, John Wiley.

Narr K. L., Thompson P. M., Szeszko P., Robinson D., Jang S., Woods R. P., Kim S., Hayashi K. M., Asunction D., Toga A. W. and Bilder R. M. (2004), "Regional Specificity of Hippocampal Volume Reductions in First-episode Schizophrenia," *NeuroImage*, 21, 1563-75.

Newey, W. K. (1993), "Efficient Estimation of Models with Conditional Moment Restrictions." In *Econometrics*, vol. 11 of *Handbook of Statistics*, 419-454, Amsterdam: North Holland.

Pizer S. M., Fletcher T., Fridman Y., Fritsch D. S., Gash A. G., Glotzer J. M., Joshi S., Thall A., Tracton G., Yushkevich P. and Chaney E. L. (2003). "Deformable M-Reps for 3D Medical Image Segmentation," *International Journal of Computer Vision*, 55, 85-106.

Presnell B., Morrison S. P. and Littell R. C. (1998), "Projected Multivariate Linear Models for Directional Data," *J. Am. Statist. Assoc.*, 93, 1068-1077.

Styner, M., Lieberman, J. A., McClure, R. K., Weinberger, D. R., Jones, D. W. and Gerig, G. (2005), "Morphometric Analysis of Lateral Ventricles in Schizophrenia and Healthy Controls Regarding Genetic and Disease-specific factors," *Proc. Natl. Acad. Sci. USA*, 102, 4872-4877.

Styner, M., Lieberman, J. A., Pantazis, D. and Gerig, G. (2004). "Boundary and Medial Shape Analysis of the Hippocampus in Schizophrenia," *Medical Image Analysis*, 8, 197-203.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Witten, D.M., Tibshirani, R. and Hastie, T. (2009), "Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis," *Biostatistics*, 10, 515-534.

Appendix: Proofs of Theorems 1 and 2

We need the following lemma throughout the proof of Theorems 1 and 2.

Lemma 1. (i) Under Assumption A1, if $f(\mathbf{z}, \boldsymbol{\beta})$ is a vector of continuous functions in $\boldsymbol{\beta}$ for any $\boldsymbol{\beta}$ in a compact set $\mathcal{B}$ and $\mathbf{z}$, then

$$\lim_{\delta \to 0} P(\sup_{\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathcal{B}, ||\boldsymbol{\beta}' - \boldsymbol{\beta}||_2 < \delta} ||f(\mathbf{z}, \boldsymbol{\beta}) - f(\mathbf{z}, \boldsymbol{\beta}')||_2 > \epsilon) = 0 \quad \forall \epsilon > 0. \tag{21}$$

(ii) In addition to the assumptions in (i), if $f(\mathbf{z}, \boldsymbol{\beta})$ also satisfies $\sup_{\boldsymbol{\beta} \in \mathcal{B}} ||f(\mathbf{z}, \boldsymbol{\beta})||_2 \leq G_1(\mathbf{z})$ and $E\{G_1(\mathbf{z})\} < \infty$, then

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}, ||\boldsymbol{\beta}' - \boldsymbol{\beta}||_2 < \delta} ||E\{f(\mathbf{z}, \boldsymbol{\beta}) - f(\mathbf{z}, \boldsymbol{\beta}')\}||_2 \to 0 \quad as \quad \delta \to 0 \tag{22}$$

$$and \quad \frac{1}{n} \sum_{i=1}^{n} [f(\mathbf{z}_i, \boldsymbol{\beta}) - E\{f(\mathbf{z}_i, \boldsymbol{\beta})\}] \quad is\ stochastically\ equicontinuous\ on \quad \mathcal{B}. \tag{23}$$

(iii) In addition to the assumptions in (ii), if $E\{G_1(\mathbf{z})^r\} < \infty$ for any $r > 1$, then

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} ||\frac{1}{n} \sum_{i=1}^{n} [f(\mathbf{z}_i, \boldsymbol{\beta}) - E\{f(\mathbf{z}_i, \boldsymbol{\beta})\}]||_2 \to 0 \tag{24}$$

in probability, as $n \to \infty$.

(iv) In addition to the assumptions in (ii), if $E\left\{\sup_{\boldsymbol{\beta} \in \mathcal{B}, ||\boldsymbol{\beta}' - \boldsymbol{\beta}||_2 < \delta} ||f(\mathbf{z}, \boldsymbol{\beta}) - f(\mathbf{z}, \boldsymbol{\beta}')||_2^2\right\} \leq C\delta^{\psi}$ for any $\delta > 0$ in a neighborhood of 0 and some constants C and ψ, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [f(\mathbf{z}_i, \boldsymbol{\beta}) - E\{f(\mathbf{z}_i, \boldsymbol{\beta})\}] \quad is\ stochastically\ equicontinuous\ on \quad \mathcal{B}. \tag{25}$$

The assumptions and result (21) of Lemma 1 (i) correspond to Jennrich's (1969) Theorem 2. The results in Lemma 1 (ii) correspond to Andrews' (1992) Lemma 3. The results in Lemma 1 (iii) correspond to Andrews' (1992) Theorem 1. The result in Lemma 1 (iv) is a special case of Andrews' (1994) Theorems 4 and 5.

Lemma 2. Let $E(\boldsymbol{\beta}, \boldsymbol{\beta}')$ be $E\{dist(\boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}), \boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}'))^2\}$. We assume that (i) $\mathcal{B}$ is a compact set; (ii) there is a point $\boldsymbol{\beta} \in \mathcal{B}$ such that $D(\boldsymbol{\beta}) < \infty$ and $\sup_{\boldsymbol{\beta}' \in \mathcal{B}} E(\boldsymbol{\beta}, \boldsymbol{\beta}') < \infty$; (iii) $E(\boldsymbol{\beta}, \boldsymbol{\beta}')$ is a continuous function in $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$. Then, $I_{\mathcal{B}}$ is an non-empty compact set.

Proof of Lemma 2. It follows from the triangle inequality that

$$\begin{aligned}\mathrm{dist}(\mathbf{m}, \boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}'))^2 &\leq \mathrm{dist}(\mathbf{m}, \boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}))^2 + \mathrm{dist}(\boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}), \boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}'))^2 \\ &+ 2\mathrm{dist}(\boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}), \boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta}'))\mathrm{dist}(\mathbf{m}, \boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\beta})).\end{aligned}$$

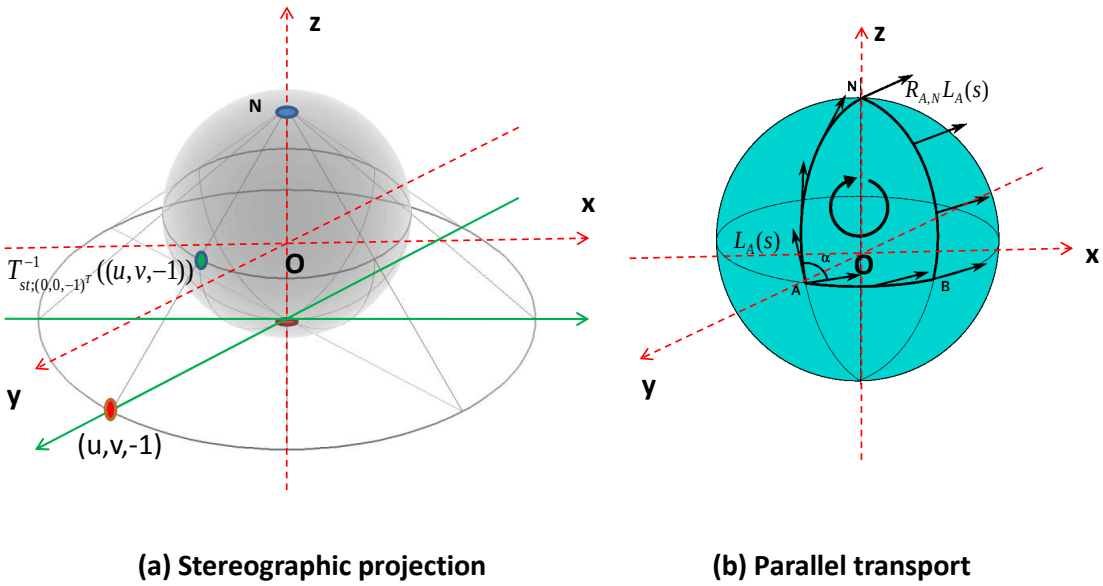
Using the Schwarz inequality and the assumptions of Lemma 2, we have

$$D(\boldsymbol{\beta}') \leq D(\boldsymbol{\beta}) + E(\boldsymbol{\beta}, \boldsymbol{\beta}') + 2\sqrt{D(\boldsymbol{\beta})E(\boldsymbol{\beta}, \boldsymbol{\beta}')} < \infty$$

for any $\boldsymbol{\beta}' \in \mathcal{B}$. Thus, $D(\boldsymbol{\beta})$ is a real continuous function of $\boldsymbol{\beta}$ in a compact set, which yields that $I_{\mathcal{B}}$ is an non-empty set. Since $\mathcal{B}$ is a compact set, it is trivial that $I_{\mathcal{B}}$ is a compact set.

Proof of Theorem 1. We prove Theorem 1 (a) in two parts. The first part proves weak consistency of $\hat{\boldsymbol{\beta}}_E$. We set $f(\mathbf{z}, \boldsymbol{\beta}) = \mathrm{dist}(\mathbf{m}, \mu(\boldsymbol{\beta}))^2 = \mathcal{E}(\boldsymbol{\beta})^T\mathcal{E}(\boldsymbol{\beta})$. It follows from Assumption A3 that $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \mathrm{dist}(\mathbf{m}, \mu(\boldsymbol{\beta}))^2 \leq G(\mathbf{z})^2$. Thus, Lemma 1 (ii) and (iii) yield that $\sup_{\boldsymbol{\beta}\in\mathcal{B}} |n^{-1}D_n(\boldsymbol{\beta}) - D(\boldsymbol{\beta})| \to 0$ in probability and $D(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$ uniformly over $\boldsymbol{\beta} \in \Theta$. Since $I_{\mathcal{B}}$ is a compact set and $\boldsymbol{\beta}_*$ is an isolated point, $\hat{\boldsymbol{\beta}}_I$ is a consistent estimator of $\boldsymbol{\beta}_*$. Furthermore, we can show that $\sup_{\boldsymbol{\beta}\in\mathcal{B}} |n^{-1}\sum_{i=1}^n [\hat{\mathbf{h}}_E(\mathbf{x}_i)\mathcal{E}_i(\boldsymbol{\beta}) - E\left\{\hat{\mathbf{h}}_E(\mathbf{x}_i)\mathcal{E}_i(\boldsymbol{\beta})\right\}]| \to 0$ in probability. Using similar arguments, we can show that $\hat{\boldsymbol{\beta}}_E$ is also a consistent estimator of $\boldsymbol{\beta}_*$. Using the results of Lemma 1, we can show the asymptotic normality of $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_I$ under conditions A1-A4 (Andrews, 1999).

Proof of Theorem 2. Using standard arguments, we can easily prove Theorem 2. Specifically, as $n \to \infty$, since it follows from Theorem 1 (ii) that $\hat{\boldsymbol{\Sigma}}_E^{-1/2}(\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_*) \to N(\mathbf{0}, \mathbf{I}_p)$, $(A\hat{\boldsymbol{\Sigma}}_E A^T)^{-1/2}A(\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_*) \to N(\mathbf{0}, \mathbf{I}_r)$, which finishes the proof of Theorem 2.



(a) Stereographic projection **(b) Parallel transport**

Figure 2: Graphic illustration of (a) stereographic projection and (b) parallel transport. In panels (a) and (b), **N** and **O** denote the north pole $(0, 0, 1)$ and the origin $(0, 0, 0)$, respectively, and the red dash lines are the x, y, and z-axes. In panel (a), the red point $(u, v, -1)$ is a selected point on the plane $z = -1$ and the green point $T^{-1}_{st;(0,0,-1)^T}((u, v, -1))$ is the inverse map of the stereographic projection mapping from $(u, v, -1)$ back to S^2. In panel (b), the point **A** is on S^2, $\mathbf{L_A}(s)$ is in $T_A S^2$, and $\mathbf{R}_{A,N}\mathbf{L}_A(s) \in T_N S^2$ is the parallel transport of $\mathbf{L}_A(s)$ from **A** to the north pole **N**.

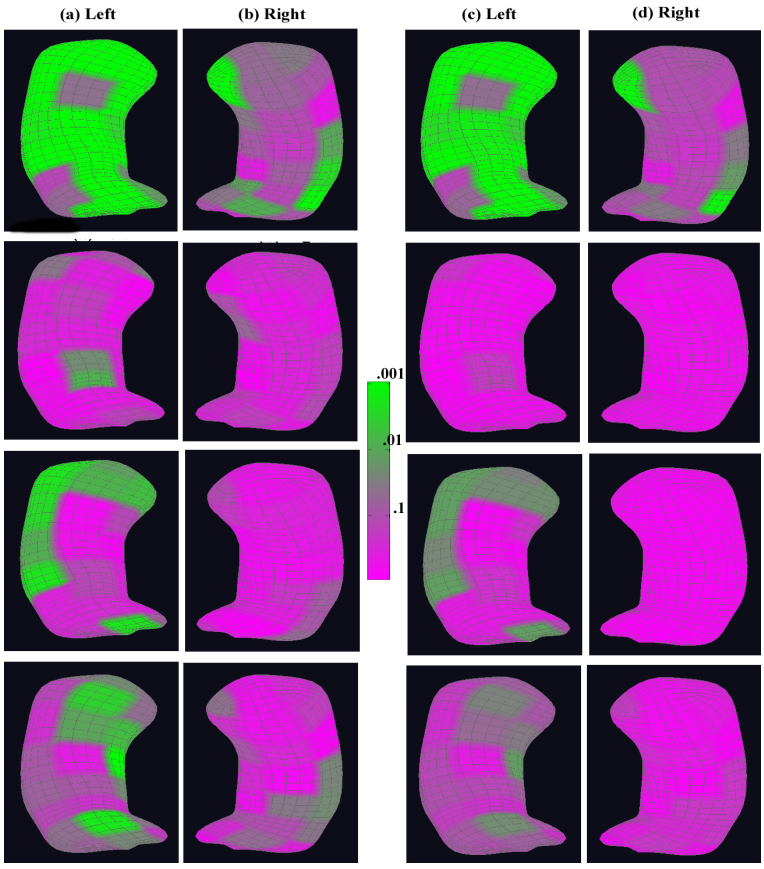


Figure 3: The coded $p-$value maps of the diagnostic status effects from the schizophrenia study of the hippocampus: rows 1, 2, 3, and 4 are for the whole medial representation structure, radius, location, and two directions, respectively: at each row, the uncorrected $p-$value maps for (a) the left hippocampus and (b) the right hippocampus; the corrected $p-$value maps for (c) the left hippocampus and (d) the right hippocampus after correcting for multiple comparisons.

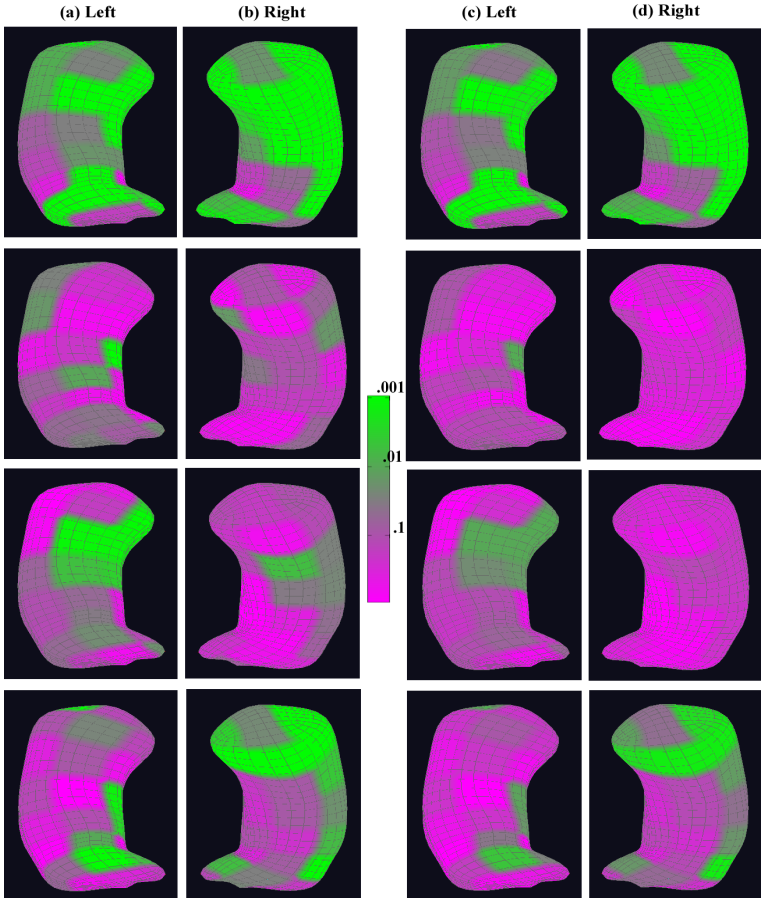


Figure 4: The color-coded $p-$value maps of the age effect from the schizophrenia study of the hippocampus: row 1, 2, 3, and 4 are for the whole medial representation structure, radius, location, and two directions, respectively: at each row, the uncorrected $p-$value maps for (a) the left hippocampus and (b) the right hippocampus; the corrected $p-$value maps for (c) the left hippocampus and (d) the right hippocampus after correcting for multiple comparisons.

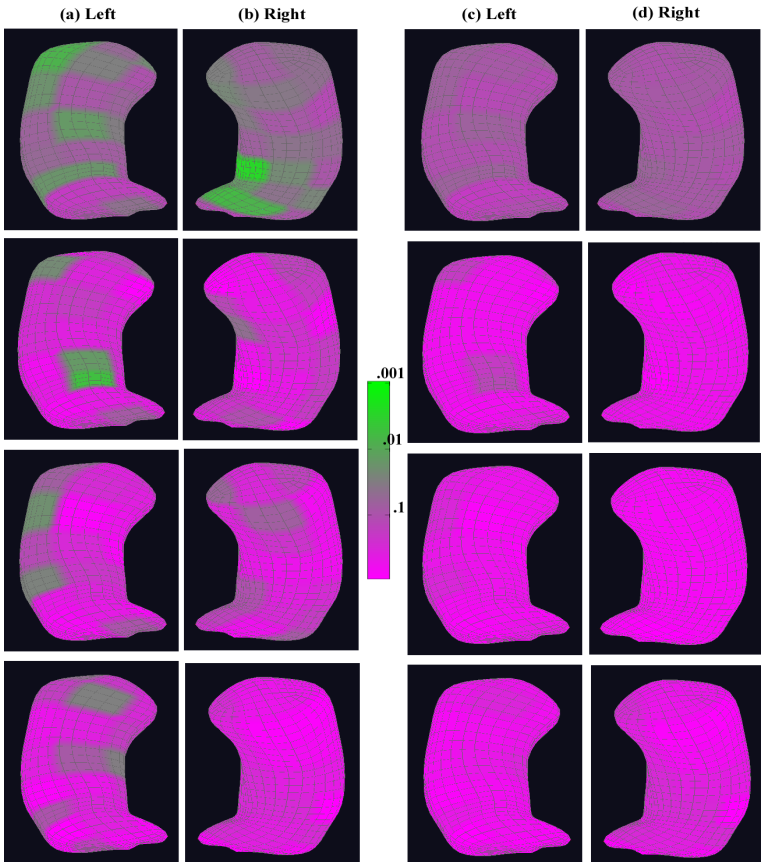


Figure 5: The coded $p-$value maps of the diagnostic status effects without accounting for other factors from the schizophrenia study of the hippocampus: rows 1, 2, 3, and 4 are for the whole medial representation structure, radius, location, and two directions, respectively: at each row, the uncorrected $p-$value maps for (a) the left hippocampus and (b) the right hippocampus; the corrected $p-$value maps for (c) the left hippocampus and (d) the right hippocampus after correcting for multiple comparisons.