Spatially Varying Coefficient Model for Neuroimaging Data with Jump Discontinuities

Hongtu Zhu*, Department of Biostatistics
and Biomedical Research Imaging Center
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
Jianqing Fan†,
Department of Oper Res and Fin. Eng
Princeton University, Princeton, NJ 08540
and Linglong Kong‡
Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, AB Canada T6G 2G1

Abstract

Motivated by recent work on studying massive imaging data in various neuroimaging studies, we propose a novel spatially varying coefficient model (SVCM) to capture the varying association between imaging measures in a three-dimensional (3D) volume (or 2D surface) with a set of covariates. Two stylized features of neuorimaging data are the presence of multiple piecewise smooth regions with unknown edges and jumps and substantial spatial correlations. To specifically account for these two features, SVCM includes a measurement model with multiple varying coefficient functions, a jumping surface model for each varying coefficient function, and a functional principal component model. We develop a three-stage estimation procedure to simultaneously estimate the varying coefficient functions and the spatial correlations. The estimation procedure includes a fast multiscale adaptive estimation and testing procedure to independently estimate each varying coefficient function, while preserving its edges among different piecewise-smooth regions. We systematically investigate the asymptotic properties (e.g., consistency and asymptotic normality) of the multiscale adaptive parameter estimates.

We also establish the uniform convergence rate of the estimated spatial covariance function and its associated eigenvalues and eigenfunctions. Our Monte Carlo simulation and real data analysis have confirmed the excellent performance of SVCM.

Key Words: Asymptotic normality; Functional principal component analysis; Jumping surface model; Kernel; Spatial varying coefficient model; Wald test.

1 Introduction

The aims of this paper are to develop a spatially varying coefficient model (SVCM) to delineate association between massive imaging data and a set of covariates of interest, such as age, and to characterize the spatial variability of the imaging data. Examples of such imaging data include T1 weighted magnetic resonance imaging (MRI), functional MRI, and diffusion tensor imaging, among many others (Friston, 2007; Thompson and Toga, 2002; Mori, 2002; Lazar, 2008). In neuroimaging studies, following spatial normalization, imaging data usually consists of data points from different subjects (or scans) at a large number of locations (called voxels) in a common 3D volume (without loss of generality), which is called a *template*. We assume that all imaging data have been registered to a template throughout the paper.

To analyze such massive imaging data, researchers face at least two main challenges. The first one is to characterize varying association between imaging data and covariates, while preserving important features, such as edges and jumps, and the shape and spatial extent of effect images. Due to the physical and biological reasons, imaging data are usually expected to contain spatially contiguous regions or effect regions with relatively sharp edges (Chumbley et al., 2009; Chan and Shen, 2005; Tabelow et al., 2008a,b). For instance, normal brain tissue can generally be classified into three broad tissue types including white matter, gray matter, and cerebrospinal fluid. These three tissues can be roughly separated by using MRI due to their imaging intensity differences and relatively intensity homogeneity within each tissue. The second challenge is to characterize spatial correlations among a large number of voxels, usually in the tens thousands to millions, for imaging data. Such spatial correlation structure and variability are important for achieving better prediction accuracy, for increasing the sensitivity of signal detection, and for characterizing the random variability of imaging data across subjects (Cressie and Wikle, 2011; Spence et al., 2007).

There are two major statistical methods including voxel-wise methods and multiscale adaptive methods for addressing the first challenge. Conventional voxel-wise approaches

involve in Gaussian smoothing imaging data, independently fitting a statistical model to imaging data at each voxel, and generating statistical maps of test statistics and p-values (Lazar, 2008; Worsley et al., 2004). As shown in Chumbley et al. (2009) and Li et al. (2011), voxel-wise methods are generally not optimal in power since it ignores the spatial information of imaging data. Moreover, the use of Gaussian smoothing can blur the image data near the edges of the spatially contiguous regions and thus introduce substantial bias in statistical results (Yue et al., 2010).

There is a great interest in the development of multiscale adaptive methods to adaptively smooth neuroimaging data, which is often characterized by a high noise level and a low signal-to-noise ratio (Tabelow et al., 2008a,b; Polzehl et al., 2010; Li et al., 2011; Qiu, 2005, 2007). Such multiscale adaptive methods not only increase signal-to-noise ratio, but also preserve important features (e.g., edge) of imaging data. For instance, in Polzehl and Spokoiny (2000, 2006), a novel propagation-separation approach was developed to adaptively and spatially smooth a single image without explicitly detecting edges. Recently, there are a few attempts to extend those adaptive smoothing methods to smoothing multiple images from a single subject (Tabelow et al., 2008a,b; Polzehl et al., 2010). In Li et al. (2011), a multiscale adaptive regression model, which integrates the propagation-separation approach and voxel-wise approach, was developed for a large class of parametric models.

There are two major statistical models, including Markov random fields and low rank models, for addressing the second challenge. The Markov random field models explicitly use the Markov property of an undirected graph to characterize spatial dependence among spatially connected voxels (Besag, 1986; Li, 2009). However, it can be restrictive to assume a specific type of spatial correlation structure, such as Markov random fields, for very large spatial data sets besides its computational complexity (Cressie and Wikle, 2011). In spatial statistics, low rank models, also called spatial random effects models, use a linear combination of 'known' spatial basis functions to approximate spatial dependence structure in a single spatial map (Cressie and Wikle, 2011). The low rank

models have a close connection with the functional principal component analysis model for characterizing spatial correlation structure in multiple images, in which spatial basis functions are directly estimated (Zipunnikov et al., 2011; Ramsay and Silverman, 2005; Hall et al., 2006).

The goal of this article is to develop SVCM and its estimation procedure to simultaneously address the two challenges discussed above. SVCM has three features: piecewise smooth, spatially correlated, and spatially adaptive, while its estimation procedure is fast, accurate and individually updated. Major contributions of the paper are as follows.

- Compared with the existing multiscale adaptive methods, SVCM first integrates a jumping surface model to delineate the piecewise smooth feature of raw and effect images and the functional principal component model to explicitly incorporate the spatial correlation structure of raw imaging data.

- A comprehensive three-stage estimation procedure is developed to adaptively and spatially improve estimation accuracy and capture spatial correlations.

- Compared with the existing methods, we use a fast and accurate estimation method to independently smooth each of effect images, while consistently estimating their standard deviation images.

- We systematically establish consistency and asymptotic distribution of the adaptive parameter estimators under two different scenarios including piecewise-smooth and piecewise-constant varying coefficient functions. In particular, we introduce several adaptive boundary conditions to delineate the relationship between the amount of jumps and the sample size. Our conditions and theoretical results differ substantially from those for the propagation-separation type methods (Polzehl and Spokoiny, 2000, 2006; Li et al., 2011).

The rest of this paper is organized as follows. In Section 2, we describe SVCM and its three-stage estimation procedure and establish the theoretical properties. In Section

3, we present a set of simulation studies with the known ground truth to examine the finite sample performance of the three-stage estimation procedure for SVCM. In Section 4, we apply the proposed methods in a real imaging dataset on attention deficit hyperactivity disorder (ADHD). In Section 5, we conclude the paper with some discussions. Technical conditions are given in Section 6. Proofs and additional results are given in a supplementary document.

2 Spatial Varying Coefficient Model with Jumping Discontinuities

2.1 Model Setup

We consider imaging measurements in a template and clinical variables (e.g., age, gender, and height) from n subjects. Let $\mathcal{D}$ represent a 3D volume and $\mathbf{d}$ and $\mathbf{d}_0$, respectively, denote a point and the center of a voxel in $\mathcal{D}$. Let $\mathcal{D}_0$ be the union of all centers $\mathbf{d}_0$ in $\mathcal{D}$ and N_D equal the number of voxels in $\mathcal{D}_0$. Without loss of generality, $\mathcal{D}$ is assumed to be a compact set in R^3. For the i-th subject, we observe an $m \times 1$ vector of imaging measures $y_i(\mathbf{d}_0)$ at $\mathbf{d}_0 \in \mathcal{D}_0$, which leads to an $mN_D \times 1$ vector of measurements across $\mathcal{D}_0$, denoted by $\mathbf{Y}_{i,\mathcal{D}_0} = \{y_i(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$. For notational simplicity, we set $m = 1$ and consider a 3D volume throughout the paper.

The proposed *spatial varying coefficient model* (SVCM) consists of three components: a measurement model, a jumping surface model, and a functional component analysis model. The measurement model characterizes the association between imaging measures and covariates and is given by

$$y_i(\mathbf{d}) = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{d}) + \eta_i(\mathbf{d}) + \epsilon_i(\mathbf{d}) \quad \text{for all} \quad i = 1, \ldots, n \quad \text{and} \quad \mathbf{d} \in \mathcal{D}, \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a $p \times 1$ vector of covariates, $\boldsymbol{\beta}(\mathbf{d}) = (\beta_1(\mathbf{d}), \ldots, \beta_p(\mathbf{d}))^T$ is a $p \times 1$ vector of coefficient functions of d, $\eta_i(\mathbf{d})$ characterizes individual image variations from $\mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{d})$, and $\epsilon_i(\mathbf{d})$ are measurement errors. Moreover, $\{\eta_i(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$

is a stochastic process indexed by $\mathbf{d} \in \mathcal{D}$ that captures the within-image dependence. We assume that they are mutually independent and $\eta_i(\mathbf{d})$ and $\epsilon_i(\mathbf{d})$ are independent and identical copies of $\mathrm{SP}(\mathbf{0}, \Sigma_\eta)$ and $\mathrm{SP}(\mathbf{0}, \Sigma_\epsilon)$, respectively, where $\mathrm{SP}(\mu, \Sigma)$ denotes a stochastic process vector with mean function $\mu(\mathbf{d})$ and covariance function $\Sigma(\mathbf{d}, \mathbf{d}')$. Moreover, $\epsilon_i(\mathbf{d})$ and $\epsilon_i(\mathbf{d}')$ are independent for $\mathbf{d} \neq \mathbf{d}'$ and thus $\Sigma_\epsilon(\mathbf{d}, \mathbf{d}') = 0$ for $\mathbf{d} \neq \mathbf{d}'$. Therefore, the covariance function of $\{\mathbf{y}_i(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$, conditioned on $\mathbf{x}_i$, is given by

$$\Sigma_y(\mathbf{d}, \mathbf{d}') = \mathrm{Cov}(\mathbf{y}_i(\mathbf{d}), \mathbf{y}_i(\mathbf{d}')) = \Sigma_\eta(\mathbf{d}, \mathbf{d}') + \Sigma_\epsilon(\mathbf{d}, \mathbf{d})\mathbf{1}(\mathbf{d} = \mathbf{d}'). \tag{2}$$

The second component of the SVCM is a jumping surface model for each of $\{\beta_j(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}_{j \leq p}$. Imaging data $\{y_i(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$ can usually be regarded as a noisy version of a piecewise-smooth function of $\mathbf{d} \in \mathcal{D}$ with jumps or edges. In many neuroimaging data, those jumps or edges often reflect the functional and/or structural changes, such as white matter and gray matter, across the brain. Therefore, the varying function $\{\beta_j(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$ in model (1) may inherit the piecewise-smooth feature from imaging data for $j = 1, \ldots, p$, but allows to have different jumps and edges. Specially, we make the following assumptions.

- (i) (Disjoint Partition) There is a finite and disjoint partition $\{\mathcal{D}_{j,l} : l = 1, \cdots, L_j\}$ of $\mathcal{D}$ such that each $\mathcal{D}_{j,l}$ is a connected region of $\mathcal{D}$ and its interior, denoted by $\mathcal{D}^o_{j,l}$, is nonempty, where L_j is a fixed, but unknown integer. See Figure 1 (a), (b), and (d) for an illustration.

- (ii) (Piecewise Smoothness) $\beta_j(\mathbf{d})$ is a smooth function of $\mathbf{d}$ within each $\mathcal{D}^o_{j,l}$ for $l = 1, \ldots, L_j$, but $\beta_j(\mathbf{d})$ is discontinuous on $\partial\mathcal{D}^{(j)} = \mathcal{D} \setminus [\cup_{l=1}^{L_j} \mathcal{D}^o_{j,l}]$, which is the union of the boundaries of all $\mathcal{D}_{j,l}$. See Figure 1 (b) for an illustration.

- (iii) (Local Patch) For any $\mathbf{d}_0 \in \mathcal{D}_0$ and $h > 0$, let $B(\mathbf{d}_0, h)$ be an open ball of $\mathbf{d}_0$ with radius h and $P_j(\mathbf{d}_0, h)$ a maximal path-connected set in $B(\mathbf{d}_0, h)$, in which $\beta_j(\mathbf{d})$ is a smooth function of $\mathbf{d}$. Assume that $P_j(\mathbf{d}_0, h)$, which will be called a *local patch*, contains an open set. See Figure 1 for a graphical illustration.

The jumping surface model can be regarded as a generalization of various models for delineating changes at unknown location (or time). See, for example, Khodadadi and Asgharian (2008) for an annotated bibliography of change point problem and regression. The disjoint partition and piecewise smoothness assumptions characterize the shape and smoothness of $\beta_j(\mathbf{d})$ in $\mathcal{D}$, whereas the local patch assumption primarily characterizes the local shape of $\beta_j(\mathbf{d})$ at each voxel $\mathbf{d}_0 \in \mathcal{D}_0$ across different scales (or radii). For $\mathbf{d}_0 \in [\cup_{l=1}^{L_j} \mathcal{D}_{j,l}^o] \cap \mathcal{D}_0$, there exists a radius $h(\mathbf{d}_0)$ such that $B(\mathbf{d}_0, h(\mathbf{d}_0)) \subset \cup_{l=1}^{L_j} \mathcal{D}_{j,l}^o$. In this case, for $h \leq h(\mathbf{d}_0)$, we have $P_j(\mathbf{d}_0, h) = B(\mathbf{d}_0, h)$ and $P_j(\mathbf{d}_0, h)^c = \emptyset$, whereas $P_j(\mathbf{d}_0, h)^c$ may not equal the empty set for large h since $B(\mathbf{d}_0, h)$ may cross different $\mathcal{D}_{j,l}^o$s. For $\mathbf{d}_0 \in \partial\mathcal{D}^{(j)} \cap \mathcal{D}_0$, $P_j(\mathbf{d}_0, h)^c \neq \emptyset$ for all $h > 0$. Since $P_j(\mathbf{d}_0, h)$ contains an open set for any $h > 0$, it eliminates the case of $\mathbf{d}_0$ being an isolated point. See Figure 1 (a) and (d) for an illustration.

The last component of the SVCM is a functional principal component analysis model for $\eta_i(\mathbf{d})$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be ordered values of the eigenvalues of the linear operator determined by Σ_η with $\sum_{l=1}^{\infty} \lambda_l < \infty$ and the $\psi_l(\mathbf{d})$s' be the corresponding orthonormal eigenfunctions (or principal components) (Li and Hsing, 2010; Hall et al., 2006). Then, Σ_η admits the spectral decomposition:

$$\Sigma_\eta(\mathbf{d}, \mathbf{d}') = \sum_{l=1}^{\infty} \lambda_l \psi_l(\mathbf{d}) \psi_l(\mathbf{d}'). \tag{3}$$

The eigenfunctions $\psi_l(\mathbf{d})$ form an orthonormal basis on the space of square-integrable functions on $\mathcal{D}$, and $\eta_i(\mathbf{d})$ admits the Karhunen-Loeve expansion as follows:

$$\eta_i(\mathbf{d}) = \sum_{l=1}^{\infty} \xi_{i,l} \psi_l(\mathbf{d}), \tag{4}$$

where $\xi_{i,l} = \int_{s \in \mathcal{D}} \eta_i(s) \psi_l(s) d\mathcal{V}(s)$ is referred to as the l-th functional principal component score of the ith subject, in which $d\mathcal{V}(s)$ denotes the Lebesgue measure. The $\xi_{i,l}$ are uncorrelated random variables with $E(\xi_{i,l}) = 0$ and $E(\xi_{i,l}\xi_{i,k}) = \lambda_l \mathbf{1}(l = k)$. If $\lambda_l \approx 0$ for $l \geq L_S + 1$, then model (1) can be approximated by

$$y_i(\mathbf{d}) \approx \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{d}) + \sum_{l=1}^{L_S} \xi_{i,l} \psi_l(\mathbf{d}) + \epsilon_i(\mathbf{d}). \tag{5}$$

In (5), since $\xi_{i,l}$ are random variables and $\psi_l(\mathbf{d})$ are 'unknown' but fixed basis functions, it can be regarded as a *varying coefficient spatial mixed effects model*. Therefore, model (5) is a mixed effects representation of model (1).

Model (5) differs significantly from other models in the existing literature. Most varying coefficient models assume some degrees of smoothness on varying coefficient functions, while they do not model the within-curve dependence (Wu et al., 1998). See Fan and Zhang (2008) for a comprehensive review of varying coefficient models. Most spatial mixed effects models in spatial statistics assume that spatial basis functions are known and regression coefficients do not vary across $\mathbf{d}$ (Cressie and Wikle, 2011). Most functional principal component analysis models focus on characterizing spatial correlation among multiple observed functions when $\mathcal{D} \in R^1$ (Zipunnikov et al., 2011; Ramsay and Silverman, 2005; Hall et al., 2006).

2.2 Three-stage Estimation Procedure

We develop a three-stage estimation procedure as follows. See Figure 2 for a schematic overview of SVCM.

- Stage (I): Calculate the least squares estimate of $\boldsymbol{\beta}(\mathbf{d}_0)$, denoted by $\hat{\boldsymbol{\beta}}(\mathbf{d}_0)$, across all voxels in $\mathcal{D}_0$, and estimate $\{\Sigma_\epsilon(\mathbf{d}_0, \mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$, $\{\Sigma_\eta(\mathbf{d}, \mathbf{d}') : (\mathbf{d}, \mathbf{d}') \in \mathcal{D}^2\}$ and its eigenvalues and eigenfunctions.

- Stage (II): Use the propagation-seperation method to adaptively and spatially smooth each component of $\hat{\boldsymbol{\beta}}(\mathbf{d}_0)$ across all $\mathbf{d}_0 \in \mathcal{D}_0$.

- Stage (III): Approximate the asymptotic covariance matrix of the final estimate of $\boldsymbol{\beta}(\mathbf{d}_0)$ and calculate test statistics across all voxels $\mathbf{d}_0 \in \mathcal{D}_0$.

This is more refined idea than the two-stage procedure proposed in Fan and Zhang (1999, 2002).

2.2.1 Stage (I)

Stage (I) consists of four steps.

Step (I.1) is to calculate the least squares estimate of $\boldsymbol{\beta}(\mathbf{d}_0)$, which equals $\hat{\boldsymbol{\beta}}(\mathbf{d}_0) = \Omega_{X,n}^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i(\mathbf{d}_0)$ across all voxels $\mathbf{d}_0 \in \mathcal{D}_0$, where $\Omega_{X,n} = \sum_{i=1}^{n} \mathbf{x}_i^{\otimes 2}$, in which $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for any vector $\mathbf{a}$. See Figure 1 (c) for a graphical illustration of $\{\hat{\boldsymbol{\beta}}(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$.

Step (I.2) is to estimate $\eta_i(\mathbf{d})$ for all $\mathbf{d} \in \mathcal{D}$. We employ the local linear regression technique to estimate all individual functions $\eta_i(\mathbf{d})$. Let $\partial_d \eta_i(\mathbf{d}) = \partial \eta_i(\mathbf{d})/\partial \mathbf{d}$, $C_i(\mathbf{d}) = (\eta_i(\mathbf{d}), h\partial_d \eta_i(\mathbf{d})^T)^T$, and $\mathbf{z}_h(\mathbf{d}_m - \mathbf{d}) = (1, (d_{m,1} - d_1)/h, (d_{m,2} - d_2)/h, (d_{m,3} - d_3)/h)^T$, where $\mathbf{d} = (d_1, d_2, d_3)^T$ and $\mathbf{d}_m = (d_{m,1}, d_{m,2}, d_{m,3})^T \in \mathcal{D}_0$. We use Taylor series expansion to expand $\eta_i(\mathbf{d}_m)$ at $\mathbf{d}$ leading to

$$\eta_i(\mathbf{d}_m) = C_i(\mathbf{d})^T \mathbf{z}_h(\mathbf{d}_m - \mathbf{d}).$$

We develop an algorithm to estimate $C_i(\mathbf{d})$ as follows. Let $K_{loc}(\cdot)$ be a univariate kernel function and $K_h(\mathbf{d}_m - \mathbf{d}) = h^{-3} \prod_{k=1}^{3} K_{loc}((d_{m,k} - d_k)/h)$ be the rescaled kernel function with a bandwidth h. For each i, we estimate $C_i(\mathbf{d})$ by minimizing the weighted least squares function given by

$$\hat{C}_i(\mathbf{d}) = \text{argmin}_{C_i(\mathbf{d})} \sum_{\mathbf{d}_m \in \mathcal{D}_0} \{r_i(\mathbf{d}_m) - C_i(\mathbf{d})^T \mathbf{z}_h(\mathbf{d}_m - \mathbf{d})\}^2 K_h(\mathbf{d}_m - \mathbf{d}).$$

where $r_i(\mathbf{d}_m) = y_i(\mathbf{d}_m) - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\mathbf{d}_m)$. It can be shown that

$$\hat{C}_i(\mathbf{d}) = \{ \sum_{\mathbf{d}_m \in \mathcal{D}_0} K_h(\mathbf{d}_m - \mathbf{d}) \mathbf{z}_h(\mathbf{d}_m - \mathbf{d})^{\otimes 2} \}^{-1} \sum_{\mathbf{d}_m \in \mathcal{D}_0} K_h(\mathbf{d}_m - \mathbf{d}) \mathbf{z}_h(\mathbf{d}_m - \mathbf{d}) r_i(\mathbf{d}_m), \quad (6)$$

Let $\hat{R}_i = (r_i(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0)$ be an $N_D \times 1$ vector of estimated residuals and notice that $\hat{\boldsymbol{\eta}}_i(\mathbf{d})$ is the first component of $C_i(\mathbf{d})$. Then, we have

$$\hat{\boldsymbol{\eta}}_i = (\hat{\eta}_i(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0) = S_i \hat{R}_i \quad \text{and} \quad \hat{\eta}_i(\mathbf{d}) = (1, 0, 0, 0) \hat{C}_i(\mathbf{d}), \quad (7)$$

where S_i is an $N_D \times N_D$ smoothing matrix (Fan and Gijbels, 1996). We pool the data from all n subjects and select the optimal bandwidth h, denoted by $\tilde{h}$, by minimizing

the generalized cross-validation (GCV) score given by

$$\mathrm{GCV}(h) = \sum_{i=1}^{n} \frac{\hat{R}_i^T (I_D - S_i)^T (I_D - S_i) \hat{R}_i}{[1 - N_D^{-1} \mathrm{tr}(S_i)]^2}, \tag{8}$$

where I_D is an $N_D \times N_D$ identity matrix. Based on $\tilde{h}$, we can use (7) to estimate $\eta_i(\mathbf{d})$ for all i.

Step (I.3) is to estimate $\Sigma_\eta(\mathbf{d}, \mathbf{d}')$ and $\Sigma_\epsilon(\mathbf{d}_0, \mathbf{d}_0)$. Let $\hat{\epsilon}_i(\mathbf{d}_0) = y_i(\mathbf{d}_0) - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\mathbf{d}_0) - \hat{\eta}_i(\mathbf{d}_0)$ be estimated residuals for $i = 1, \ldots, n$ and $\mathbf{d}_0 \in \mathcal{D}_0$. We estimate $\Sigma_\epsilon(\mathbf{d}_0, \mathbf{d}_0)$ by

$$\hat{\Sigma}_\epsilon(\mathbf{d}_0, \mathbf{d}_0) = n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i(\mathbf{d}_0)^2 \tag{9}$$

and $\Sigma_\eta(\mathbf{d}, \mathbf{d}')$ by the sample covariance matrix:

$$\hat{\Sigma}_\eta(\mathbf{d}, \mathbf{d}') = (n - p)^{-1} \sum_{i=1}^{n} \hat{\eta}_i(\mathbf{d}) \hat{\eta}_i(\mathbf{d}'). \tag{10}$$

Step (I.4) is to estimate the eigenvalue-eigenfunction pairs of Σ_η by using the singular value decomposition. Let $\mathbf{V} = [\hat{\boldsymbol{\eta}}_1, \cdots, \hat{\boldsymbol{\eta}}_n]$ be an $N_D \times n$ matrix. Since n is much smaller than N_D, we can easily calculate the eigenvalue-eigenvector pairs of the $n \times n$ matrix $\mathbf{V}^T\mathbf{V}$, denoted by $\{(\hat{\lambda}_i, \hat{\boldsymbol{\xi}}_i) : i = 1, \cdots, n\}$. It can be shown that $\{(\hat{\lambda}_i, \mathbf{V}\hat{\boldsymbol{\xi}}_i) : i = 1, \cdots, n\}$ are the eigenvalue-eigenvector pairs of the $N_D \times N_D$ matrix $\mathbf{V}\mathbf{V}^T$. In applications, one usually considers large $\hat{\lambda}_l$ values, while dropping small $\hat{\lambda}_l$s. It is common to choose a value of L_S so that the cumulative eigenvalue $\sum_{l=1}^{L_S} \hat{\lambda}_l / \sum_{l=1}^{n} \hat{\lambda}_l$ is above a prefixed threshold, say 80% (Zipunnikov et al., 2011; Li and Hsing, 2010; Hall et al., 2006). Furthermore, the lth SPCA scores can be computed using

$$\hat{\xi}_{i,l} = \sum_{m=1}^{N_D} \hat{\eta}_i(\mathbf{d}_m) \hat{\psi}_l(\mathbf{d}_m) \mathcal{V}(\mathbf{d}_m) \tag{11}$$

for $l = 1, \ldots, L_S$, where $\mathcal{V}(\mathbf{d}_m)$ is the volume of voxel $\mathbf{d}_m$.

2.2.2 Stage (II)

Stage (II) is a multiscale adaptive and sequential smoothing (MASS) method. The key idea of MASS is to use the propagation-separation method (Polzehl and Spokoiny, 2000,

2006) to individually smooth each least squares estimate image $\{\hat{\beta}_j(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$ for $j = 1, \ldots, p$. MASS starts with building a sequence of nested spheres with increasing bandwidths $0 = h_0 < h_1 < \cdots < h_S = r_0$ ranging from the smallest bandwidth h_1 to the largest bandwidth $h_S = r_0$ for each $\mathbf{d}_0 \in \mathcal{D}_0$. At bandwidth h_1, based on the information contained in $\{\hat{\boldsymbol{\beta}}(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$, we sequentially calculate adaptive weights $\omega_j(\mathbf{d}_0, \mathbf{d}_0'; h_1)$ between voxels $\mathbf{d}_0$ and $\mathbf{d}_0'$, which depends on the distance $\|\mathbf{d}_0 - \mathbf{d}_0\|$ and spacial similarity $|\hat{\beta}_j(\mathbf{d}_0) - \hat{\beta}_j(\mathbf{d}_0)|$, and update $\hat{\beta}_j(\mathbf{d}_0; h_1)$ for all $\mathbf{d}_0 \in \mathcal{D}_0$ for $j = 1, \cdots, p$. At bandwidth h_2, we repeat the same process using $\{\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_1) : \mathbf{d}_0 \in \mathcal{D}_0\}$ to compute spatial similarities. In this way, we can sequentially determine $\omega_j(\mathbf{d}_0, \mathbf{d}_0'; h_s)$ and $\hat{\beta}_j(\mathbf{d}_0; h_s)$ for each component of $\boldsymbol{\beta}(\mathbf{d}_0)$ as the bandwidth ranges from h_1 to $h_S = r_0$. Moreover, as shown below, we have found a simple way of calculating the standard deviation of $\hat{\beta}_j(\mathbf{d}_0; h_s)$.

MASS consists of three steps including (II.1) an initialization step, (II.2) a sequentially adaptive estimation step, and (II.3) a stop checking step, each of which involves in the specification of several parameters. Since propagation-separation and the choice of their associated parameters have been discussed in details in Polzehl et al. (2010) and Li et al. (2011), we briefly mention them here for the completeness. In the initialization step (II.1), we take a geometric series $\{h_s = c_h^s : s = 1, \ldots, S\}$ of radii with $h_0 = 0$, where $c_h > 1$, say $c_h = 1.10$. We suggest relatively small c_h to prevent incorporating too many neighboring voxels.

In the sequentially adaptive estimation step (II.2), starting from $s = 1$ and $h_1 = c_h$, at step s, we compute spatial adaptive locally weighted average estimate $\hat{\beta}_j(\mathbf{d}_0; h_s)$ based on $\{\hat{\beta}_j(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$ and $\{\hat{\beta}_j(\mathbf{d}_0; h_{s-1}) : \mathbf{d} \in \mathcal{D}_0\}$, where $\hat{\beta}_j(\mathbf{d}_0; h_0) = \hat{\beta}_j(\mathbf{d}_0)$. Specifically, for each j, we construct a weighted quadratic function

$$\ell_n(\beta_j(\mathbf{d}_0); h_s) = \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \{\hat{\beta}_j(\mathbf{d}_m) - \beta_j(\mathbf{d}_0)\}^2 \omega_j(\mathbf{d}_0, \mathbf{d}_m; h_s), \tag{12}$$

where $\omega_j(\mathbf{d}_0, \mathbf{d}_m; h_s)$, which will be defined below, characterizes the similarity between

$\hat{\beta}_j(\mathbf{d}_m; h_{s-1})$ and $\hat{\beta}_j(\mathbf{d}_0; h_{s-1})$. We then calculate

$$\hat{\beta}_j(\mathbf{d}_0; h_s) = \text{argmin}_{\beta_j(\mathbf{d}_0)} \ell_n(\beta_j(\mathbf{d}_0); h_s) = \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \tilde{\omega}_j(\mathbf{d}_0, \mathbf{d}_m; h_s) \hat{\beta}_j(\mathbf{d}_m), \quad (13)$$

where $\tilde{\omega}_j(\mathbf{d}_0, \mathbf{d}_m; h_s) = \omega_j(\mathbf{d}_0, \mathbf{d}_m; h_s) / \sum_{\mathbf{d}_{m'} \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \omega_j(\mathbf{d}_0, \mathbf{d}_{m'}; h_s)$.

Let $\Sigma_n(\hat{\beta}_j(\mathbf{d}_0; h_s))$ be the asymptotic variance of $\hat{\beta}_j(\mathbf{d}_0; h_s)$. For $\beta_j(\mathbf{d}_0)$, we compute the similarity between voxels $\mathbf{d}_0$ and $\mathbf{d}_0'$, denoted by $D_{\beta_j}(\mathbf{d}_0, \mathbf{d}_0'; h_{s-1})$, and the adaptive weight $\omega_j(\mathbf{d}_0, \mathbf{d}_0'; h_s)$, which are, respectively, defined as

$$\begin{aligned} D_{\beta_j}(\mathbf{d}_0, \mathbf{d}_0'; h_{s-1}) &= \{\hat{\beta}_j(\mathbf{d}_0; h_{s-1}) - \hat{\beta}_j(\mathbf{d}_0'; h_{s-1})\}^2 / \Sigma_n(\hat{\beta}_j(\mathbf{d}_0; h_{s-1})), \quad (14) \\ \omega_j(\mathbf{d}_0, \mathbf{d}_0'; h_s) &= K_{loc}(||\mathbf{d}_0 - \mathbf{d}_0'||_2 / h_s) K_{st}(D_{\beta_j}(\mathbf{d}_0, \mathbf{d}_0'; h_{s-1}) / C_n), \end{aligned}$$

where $K_{st}(u)$ is a nonnegative kernel function with compact support, C_n is a tuning parameter depending on n, and $||\cdot||_2$ denotes the Euclidean norm of a vector.

The weights $K_{loc}(||\mathbf{d}_0 - \mathbf{d}_0'||_2 / h_s)$ give less weight to the voxel $\mathbf{d}_0'$ that is far from the voxel $\mathbf{d}_0$. The weights $K_{st}(u)$ downweight the voxels $\mathbf{d}_0'$ with large $D_{\beta_j}(\mathbf{d}_0, \mathbf{d}_0'; h_{s-1})$, which indicates a large difference between $\hat{\beta}_j(\mathbf{d}_0'; h_{s-1})$ and $\hat{\beta}_j(\mathbf{d}_0; h_{s-1})$. In practice, we set $K_{loc}(u) = (1-u)_+$. Although different choices of $K_{st}(\cdot)$ have been suggested in the propagation-separation method (Polzehl and Spokoiny, 2000, 2006; Polzehl et al., 2010; Li et al., 2011), we have tested these kernel functions and found that $K_{st}(u) = \exp(-u)$ performs reasonably well. Another good choice of $K_{st}(u)$ is $\min(1, 2(1-u))_+$. Moreover, theoretically, as shown in Scott (1992) and Fan (1993), they have examined the efficiency of different kernels for weighted least squares estimators, but extending their results to the propagation-separation method needs some further investigation.

The scale C_n is used to penalize the similarity between any two voxels $\mathbf{d}_0$ and $\mathbf{d}_0'$ in a similar manner to bandwidth, and an appropriate choice of C_n is crucial for the behavior of the propagation-separation method. As discussed in (Polzehl and Spokoiny, 2000, 2006), a propagation condition independent of the observations at hand can be used to specify C_n. The basic idea of the propagation condition is that the impact of the statistical penalty in $K_{st}(D_{\beta_j}(\mathbf{d}_0, \mathbf{d}_0'; h_{s-1}) / C_n)$ should be negligible under a homogeneous

model $\beta_j(\mathbf{d}) \equiv$ constant yielding almost free smoothing within homogeneous regions. However, we take an alternative approach to choose C_n here. Specifically, a good choice of C_n should balance between the sensitivity and specificity of MASS. Theoretically, as shown in Section 2.3, C_n should satisfy $C_n/n = o(1)$ and $C_n^{-1}\log(N_D) = o(1)$. We choose $C_n = n^{0.4}\chi_1^2(0.8)$ based on our experiments, where $\chi_1^2(a)$ is the upper a-percentile of the χ_1^2-distribution.

We now calculate $\Sigma_n(\hat{\beta}_j(\mathbf{d}_0; h_s))$. By treating the weights $\tilde{\omega}_j(\mathbf{d}_0, \mathbf{d}_m; h_s)$ as 'fixed' constants, we can approximate $\Sigma_n(\hat{\beta}_j(\mathbf{d}_0; h_s))$ by

$$\sum_{\mathbf{d}_m, \mathbf{d}_{m'} \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \tilde{\omega}_j(\mathbf{d}_0, \mathbf{d}_m; h_s)\tilde{\omega}_j(\mathbf{d}_0, \mathbf{d}_{m'}; h_s)\mathrm{Cov}(\hat{\beta}_j(\mathbf{d}_m), \hat{\beta}_j(\mathbf{d}_{m'})), \tag{15}$$

where $\mathrm{Cov}(\hat{\beta}_j(\mathbf{d}_m), \hat{\beta}_j(\mathbf{d}_{m'}))$ can be estimated by

$$\mathbf{e}_{j,p}^T \Omega_{X,n}^{-1} \mathbf{e}_{j,p} \{\hat{\Sigma}_\eta(\mathbf{d}_m, \mathbf{d}_{m'}) + \hat{\Sigma}_\epsilon(\mathbf{d}_m, \mathbf{d}_m)\mathbf{1}(\mathbf{d}_m = \mathbf{d}_{m'})\}, \tag{16}$$

in which $\mathbf{e}_{j,p}$ is a $p \times 1$ vector with the j-th element 1 and others 0. We will examine the consistency of approximation (15) later.

In the stop checking step (II.3), after the first iteration, we start to calculate a stopping criterion based on a normalized distance between $\hat{\beta}_j(\mathbf{d}_0)$ and $\hat{\beta}_j(\mathbf{d}_0; h_s)$ given by

$$D(\hat{\beta}_j(\mathbf{d}_0), \hat{\beta}_j(\mathbf{d}_0; h_s)) = \{\hat{\beta}_j(\mathbf{d}_0) - \hat{\beta}_j(\mathbf{d}_0; h_s)\}^2 / \Sigma_n(\hat{\beta}_j(\mathbf{d}_0)). \tag{17}$$

Then, we check whether $\hat{\beta}_j(\mathbf{d}_0; h_s)$ is in a confidence ellipsoid of $\hat{\beta}_j(\mathbf{d}_0)$ given by $\{\beta_j(\mathbf{d}_0) : D(\hat{\beta}_j(\mathbf{d}_0), \beta_j(\mathbf{d}_0)) \le C_s\}$, where C_s is taken as $C_s = \chi_1^2(0.80/s)$ in our implementation. If $D(\hat{\beta}_j(\mathbf{d}_0), \hat{\beta}_j(\mathbf{d}_0; h_s))$ is greater than C_s, then we set $\hat{\beta}_j(\mathbf{d}_0, h_S) = \hat{\beta}_j(\mathbf{d}_0, h_{s-1})$ and $s = S$ for the j-th component and voxel $\mathbf{d}_0$. If $s = S$ for all components in all voxels, we stop. If $D(\hat{\beta}_j(\mathbf{d}_0), \hat{\beta}_j(\mathbf{d}_0; h_s)) \le C_s$, then we set $h_{s+1} = c_h h_s$, increase s by 1 and continue with the step (II.1). It should be noted that different components of $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h)$ may stop at different bandwidths.

We usually set the maximal step S to be relatively small, say between 10 and 20, and thus each $B(\mathbf{d}_0, h_S)$ only contains a relatively small number of voxels. As S increases,

the number of neighboring voxels in $B(\mathbf{d}_0, h_S)$ increases exponentially. It increases the chance of oversmoothing $\beta_j(\mathbf{d}_0)$ when $\mathbf{d}_0$ is near the edge of distinct regions. Moreover, in order to prevent oversmoothing $\beta_j(\mathbf{d}_0)$, we compare $\hat{\beta}_j(\mathbf{d}_0; h_s)$ with the least squares estimate $\hat{\beta}_j(\mathbf{d}_0)$ and gradually decrease C_s with the number of iteration.

2.2.3 Stage (III)

Based on $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S)$, we can further construct test statistics to examine scientific questions associated with $\boldsymbol{\beta}(\mathbf{d}_0)$. For instance, such questions may compare brain structure across different groups (normal controls versus patients) or detect change in brain structure across time. These questions can be formulated as the linear hypotheses about $\boldsymbol{\beta}(\mathbf{d}_0)$ given by

$$H_0(\mathbf{d}_0) : R_1\boldsymbol{\beta}(\mathbf{d}_0) = \mathbf{b}_0 \quad \text{vs.} \quad H_1(\mathbf{d}_0) : R_1\boldsymbol{\beta}(\mathbf{d}_0) \neq \mathbf{b}_0, \tag{18}$$

where R_1 is an $r \times k$ matrix of full row rank and $\mathbf{b}_0$ is an $r \times 1$ specified vector. We use the Wald test statistic

$$W_\beta(\mathbf{d}_0; h) = \{R_1\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S) - \mathbf{b}_0\}^T\{R_1\Sigma_n(\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S))R_1^T\}^{-1}\{R_1\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S) - \mathbf{b}_0\} \tag{19}$$

for problem (18), where $\Sigma_n(\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S))$ is the covariance matrix of $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S)$.

We propose an approximation of $\Sigma_n(\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S))$. According to (13), we know that

$$\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S) = \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_S)} \tilde{\boldsymbol{\omega}}(\mathbf{d}_0, \mathbf{d}_m; h_S) \circ \hat{\boldsymbol{\beta}}(\mathbf{d}_m)$$

where $\boldsymbol{a} \circ \boldsymbol{b}$ denotes the Hadamard product of matrices $\boldsymbol{a}$ and $\boldsymbol{b}$ and $\tilde{\boldsymbol{\omega}}(\mathbf{d}_0, \mathbf{d}_m; h)$ is a $p \times 1$ vector determined by the weights $\tilde{\omega}_j(\mathbf{d}_0, \mathbf{d}_m; h)$ in Stage II. Let J_p be the $p^2 \times p$ selection matrix (Liu, 1999). Therefore, $\Sigma_n(\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_S))$ can be approximated by

$$\begin{aligned} & \sum_{\mathbf{d}_m, \mathbf{d}'_m \in B(\mathbf{d}_0, h_S)} \mathrm{Cov}(\tilde{\boldsymbol{\omega}}(\mathbf{d}_0, \mathbf{d}_m; h_S) \circ \hat{\boldsymbol{\beta}}(\mathbf{d}_m), \tilde{\boldsymbol{\omega}}(\mathbf{d}_0, \mathbf{d}'_m; h_S) \circ \hat{\boldsymbol{\beta}}(\mathbf{d}'_m)) \\ \approx & \sum_{\mathbf{d}_m, \mathbf{d}'_m \in B(\mathbf{d}_0, h_S)} \hat{\Sigma}_y(\mathbf{d}_m, \mathbf{d}'_m) J_p^T\{[\tilde{\boldsymbol{\omega}}(\mathbf{d}_0, \mathbf{d}_m; h_S)\tilde{\boldsymbol{\omega}}(\mathbf{d}_0, \mathbf{d}'_m; h_S)^T] \otimes \Omega_{X,n}^{-1}\} J_p. \end{aligned}$$

2.3 Theoretical Results

We systematically investigate the asymptotic properties of all estimators obtained from the three-stage estimation procedure. Throughout the paper, we only consider a finite number of iterations and bounded r_0 for MASS, since a brain volume is always bounded. Without otherwise stated, we assume that $o_p(1)$ and $O_p(1)$ hold uniformly across all $\mathbf{d}$ in either $\mathcal{D}$ or $\mathcal{D}_0$ throughout the paper. Moreover, the sample size n and the number of voxels N_D are allowed to diverge to infinity. We state the following theorems, whose detailed assumptions and proofs can be found in Section 6 and a supplementary document.

Let $\boldsymbol{\beta}_*(\mathbf{d}_0) = (\beta_{1*}(\mathbf{d}_0), \ldots, \beta_{p*}(\mathbf{d}_0))^T$ be the true value of $\boldsymbol{\beta}(\mathbf{d}_0)$ at voxel $\mathbf{d}_0$. We first establish the uniform convergence rate of $\{\hat{\boldsymbol{\beta}}(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$.

Theorem 1. *Under assumptions (C1)-(C4) in Section 6, as $n \to \infty$, we have*

- *(i)* $\sqrt{n}[\hat{\boldsymbol{\beta}}(\mathbf{d}_0) - \boldsymbol{\beta}_*(\mathbf{d}_0)] \to^L N(\mathbf{0}, \Omega_X^{-1}\Sigma_y(\mathbf{d}_0, \mathbf{d}_0))$ *for any* $\mathbf{d}_0 \in \mathcal{D}_0$, *where* $\to^L$ *denotes convergence in distribution;*
- *(ii)* $\sup_{\mathbf{d}_0 \in \mathcal{D}_0} ||\hat{\boldsymbol{\beta}}(\mathbf{d}_0) - \boldsymbol{\beta}_*(\mathbf{d}_0)||_2 = O_p(\sqrt{n^{-1}\log(1+N_D)})$

Remark 1. Theorem 1 (i) just restates a standard asymptotic normality of the least squares estimate of $\boldsymbol{\beta}(\mathbf{d}_0)$ at any given voxel $\mathbf{d}_0 \in \mathcal{D}_0$. Theorem 1 (ii) states that the maximum of $||\hat{\boldsymbol{\beta}}(\mathbf{d}_0) - \boldsymbol{\beta}_*(\mathbf{d}_0)||_2$ across all $\mathbf{d}_0 \in \mathcal{D}_0$ is at the order of $\sqrt{n^{-1}\log(1+N_D)}$. If $\log(1+N_D)$ is relatively small compared with n, then the estimation errors converge uniformly to zero in probability. In practice, N_D is determined by imaging resolution and its value can be much larger than the sample size. For instance, in most applications, N_D can be as large as 100^3 and $\log(1+N_D)$ is around 15. In a study with several hundreds subjects, $n^{-1}\log(1+N_D)$ can be relatively small.

We next study the uniform convergence rate of $\hat{\Sigma}_\eta$ and its associated eigenvalues and eigenfunctions. We also establish the uniform convergence of $\hat{\Sigma}_\epsilon(\mathbf{d}_0, \mathbf{d}_0)$.

Theorem 2. *Under assumptions (C1)-(C8) in Section 6, we have the following results:*

$$(i) \quad \sup_{(\mathbf{d},\mathbf{d}')\in\mathcal{D}^2} |\hat{\Sigma}_\eta(\mathbf{d},\mathbf{d}') - \Sigma_\eta(\mathbf{d},\mathbf{d}')| = o_p(1);$$

$$(ii) \quad \int_{\mathcal{D}} [\hat{\psi}_l(\mathbf{d}) - \psi_l(\mathbf{d})]^2 d\mathcal{V}(\mathbf{d}) = o_p(1) \text{ and } |\hat{\lambda}_l - \lambda_l| = o_p(1) \quad \text{for } l = 1, \ldots, E;$$

$$(iii) \quad \sup_{\mathbf{d}_0\in\mathcal{D}_0} |\hat{\Sigma}_\epsilon(\mathbf{d}_0,\mathbf{d}_0) - \Sigma_\epsilon(\mathbf{d}_0,\mathbf{d}_0)| = o_p(1);$$

where E will be described in assumption (C8) and $\hat{\psi}_l(\mathbf{d})$ is the estimated eigenvector, computed from $\hat{\psi}_l = \mathbf{V}\boldsymbol{\xi}_l$.

REMARK 2. Theorem 2 (i) and (ii) characterize the uniform weak convergence of $\hat{\Sigma}_\eta(\cdot,\cdot)$ and the convergence of $\hat{\psi}_l(\cdot)$ and $\hat{\lambda}_l$. These results can be regarded as an extension of Theorems 3.3-3.6 in Li and Hsing (2010), which established the uniform strong convergence rates of these estimates under a simple model. Specifically, in Li and Hsing (2010), they considered $y_i(\mathbf{d}) = \mu(\mathbf{d}) + \eta_i(\mathbf{d}) + \epsilon_i(\mathbf{d})$ and assumed that $\mu(\mathbf{d})$ is twice differentiable. Another key difference is that in Li and Hsing (2010), they employed all cross products $y_i(\mathbf{d})y_i(\mathbf{d}')$ for $\mathbf{d} \neq \mathbf{d}'$ and then used the local polynomial kernel to estimate $\Sigma_\eta(\mathbf{d},\mathbf{d}')$. In contrast, our approach is computationally simple and $\hat{\Sigma}_\eta(\mathbf{d},\mathbf{d}')$ is positive definite. Theorem 2 (iii) characterizes the uniform weak convergence of $\hat{\Sigma}_\epsilon(\mathbf{d}_0,\mathbf{d}_0)$ across all voxels $\mathbf{d}_0 \in \mathcal{D}_0$.

To investigate the asymptotic properties of $\hat{\beta}_j(\mathbf{d}_0; h_s)$, we need to characterize points close to and far from the boundary set $\partial\mathcal{D}^{(j)}$. For a given bandwidth h_s, we first define h_s-boundary sets:

$$\partial\mathcal{D}^{(j)}(h_s) = \{\mathbf{d} \in \mathcal{D} : B(\mathbf{d}, h_s) \cap \partial\mathcal{D}^{(j)} \neq \emptyset\} \text{ and } \partial\mathcal{D}_0^{(j)}(h_s) = \partial\mathcal{D}^{(j)}(h_s) \cap \mathcal{D}_0. \quad (20)$$

Thus, $\partial\mathcal{D}^{(j)}(h_s)$ can be regarded as a band with radius h_s covering the boundary set $\partial\mathcal{D}^{(j)}$, while $\partial\mathcal{D}_0^{(j)}(h_s)$ contains all grid points within such band. It is easy to show that for a sequence of bandwidths $h_0 = 0 < h_1 < \cdots < h_S$, we have

$$\partial\mathcal{D}^{(j)}(h_0) = \partial\mathcal{D}^{(j)} \subset \cdots \subset \partial\mathcal{D}^{(j)}(h_S) \quad \text{and} \quad \partial\mathcal{D}_0^{(j)}(h_0) \subset \cdots \subset \partial\mathcal{D}_0^{(j)}(h_S). \quad (21)$$

Therefore, for a fixed bandwidth h_s, any point $\mathbf{d}_0 \in \mathcal{D}_0$ belongs to either $\mathcal{D} \setminus \partial\mathcal{D}^{(j)}(h_s)$

or $\partial\mathcal{D}^{(j)}(h_s)$. For each $\mathbf{d}_0 \in \mathcal{D} \setminus \partial\mathcal{D}^{(j)}(h_s)$, there exists one and only one $\mathcal{D}_{j,l}$ such that

$$B(\mathbf{d}_0, h_0) \subset \cdots \subset B(\mathbf{d}_0, h_s) \subset \mathcal{D}^o_{j,l}. \tag{22}$$

See Figure 1 (d) for an illustration.

We first investigate the asymptotic behavior of $\hat{\beta}_j(\mathbf{d}_0; h_s)$ when $\beta_{j*}(\mathbf{d})$ is piecewise constant. That is, $\beta_{j*}(\mathbf{d})$ is a constant in $\mathcal{D}^o_{j,l}$ and for any $\mathbf{d}' \in \partial\mathcal{D}^{(j)}$, there exists a $\mathbf{d} \in \cup_{l=1}^{L_j} \mathcal{D}^o_{j,l}$ such that $\beta_{j*}(\mathbf{d}) = \beta_{j*}(\mathbf{d}')$. Let $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s) = \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \tilde{\omega}_j(\mathbf{d}_0, \mathbf{d}_m; h_s) \beta_{j*}(\mathbf{d}_m)$ be the pseudo-true value of $\beta_j(\mathbf{d}_0)$ at scale h_s in voxel $\mathbf{d}_0$. For all $\mathbf{d}_0 \in \mathcal{D} \setminus \partial\mathcal{D}^{(j)}(h_S)$, we have $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s) = \beta_{j*}(\mathbf{d}_0)$ for all $s \leq S$ due to (22). In contrast, for $\mathbf{d}_0 \in \partial\mathcal{D}^{(j)}(h_S)$, $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s)$ may vary from h_0 to h_S. In this case, we are able to establish several important theoretical results to characterize the asymptotic behavior of $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_s)$ even when h_S does not converge to zero. We need additional notation as follows:

$$\begin{aligned}
&\hat{\Delta}_j(\mathbf{d}_0) = \hat{\beta}_j(\mathbf{d}_0) - \beta_{j*}(\mathbf{d}_0) \quad \text{and} \quad \Delta_{j*}(\mathbf{d}_0, \mathbf{d}_0') = \beta_{j*}(\mathbf{d}_0) - \beta_{j*}(\mathbf{d}_0'), \\
&\omega_j^{(0)}(\mathbf{d}_0, \mathbf{d}_0'; h_s) = K_{loc}(||\mathbf{d}_0 - \mathbf{d}_0'||_2/h_s) K_{st}(0) \mathbf{1}(\beta_{j*}(\mathbf{d}_0) = \beta_{j*}(\mathbf{d}_0')), \\
&\tilde{\omega}_j^{(0)}(\mathbf{d}_0, \mathbf{d}_0'; h_s) = \omega_j^{(0)}(\mathbf{d}_0, \mathbf{d}_0'; h_s) / \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \omega_j^{(0)}(\mathbf{d}_0, \mathbf{d}_m; h_s), \\
&\Sigma_j^{(0)}(\mathbf{d}_0; h_s) = \mathbf{e}_{j,p}^T \Omega_X^{-1} \mathbf{e}_{j,p} \sum_{\mathbf{d}_m, \mathbf{d}_m' \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \tilde{\omega}_j^{(0)}(\mathbf{d}_0, \mathbf{d}_m; h_s) \tilde{\omega}_j^{(0)}(\mathbf{d}_0, \mathbf{d}_m'; h_s) \Sigma_y(\mathbf{d}_m, \mathbf{d}_m').
\end{aligned} \tag{23}$$

Theorem 3. *Under assumptions (C1)-(C10) in Section 6 for piecewise constant* $\{\beta_{j*}(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$*, we have the following results for all* $0 \leq s \leq S$*:*

(i) $\sup_{\mathbf{d}_0 \in \mathcal{D}_0} |\tilde{\beta}_{j*}(\mathbf{d}_0; h_s) - \beta_{j*}(\mathbf{d}_0)| = o_p(\sqrt{\log(1 + N_D)/n})$*;*

(ii) $\hat{\beta}_j(\mathbf{d}_0; h_s) - \beta_{j*}(\mathbf{d}_0) = \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \tilde{\omega}_j^{(0)}(\mathbf{d}_0, \mathbf{d}_m; h_s) \hat{\Delta}_j(\mathbf{d}_m)[1 + o_p(1)]$*;*

(iii) $\sup_{\mathbf{d}_0 \in \mathcal{D}_0} |\hat{\Sigma}(\sqrt{n}\tilde{\beta}_{j*}(\mathbf{d}_0; h_s)) - \Sigma_j^{(0)}(\mathbf{d}_0; h_s)| = o_p(1)$*;*

(iv) $\sqrt{n}[\hat{\beta}_j(\mathbf{d}_0; h_s) - \beta_{j*}(\mathbf{d}_0)]$ *converges in distribution to a normal distribution with mean zero and variance* $\Sigma_j^{(0)}(\mathbf{d}_0; h_s)$ *as* $n \to \infty$*.*

Remark 3. Theorem 3 shows that MASS has several important features for a piecewise constant function $\beta_{j*}(\mathbf{d})$. For instance, Theorem 3 (i) quantifies the maximum absolute difference (or bias) between the true value $\beta_{j*}(\mathbf{d}_0)$ and the pseudo true value $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s)$ across all $\mathbf{d}_0 \in \mathcal{D}_0$ for any s. Since $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s) - \beta_{j*}(\mathbf{d}_0) = 0$ for $\mathbf{d}_0 \in \mathcal{D} \setminus \partial\mathcal{D}^{(j)}(h_s)$, this result delineates the potential bias for voxels $\mathbf{d}_0$ in $\partial\mathcal{D}^{(j)}(h_s)$.

Theorem 3 (iv) ensures that $\sqrt{n}[\hat{\beta}_j(\mathbf{d}_0; h_s) - \beta_{j*}(\mathbf{d}_0)]$ is asymptotically normally distributed. Moreover, as shown in the supplementary document, $\Sigma_j^{(0)}(\mathbf{d}_0; h_s)$ is smaller than the asymptotic variance of the raw estimate $\hat{\beta}_j(\mathbf{d}_0)$. As a result, MASS increases statistical power of testing $H_0(\mathbf{d}_0)$.

We now consider a much complex scenario when $\beta_{j*}(\mathbf{d})$ is piecewise smooth. In this case, $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s)$ may vary from h_0 to h_S for all voxels $\mathbf{d}_0 \in \mathcal{D}_0$ regardless whether $\mathbf{d}_0$ belongs to $\partial\mathcal{D}^{(j)}(h_s)$ or not. We can establish important theoretical results to characterize the asymptotic behavior of $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_s)$ only when $h_s = O(\sqrt{\log(1+N_D)/n}) = o(1)$ holds. We need some additional notation as follows:

$$\omega_j^{(1)}(\mathbf{d}_0, \mathbf{d}_0'; h_s) = K_{loc}(||\mathbf{d}_0 - \mathbf{d}_0'||_2/h_s) K_{st}(0) \mathbf{1}(|\beta_{j*}(\mathbf{d}_0) - \beta_{j*}(\mathbf{d}_0')| \le O(h_s)), \tag{24}$$

$$\tilde{\omega}_j^{(1)}(\mathbf{d}_0, \mathbf{d}_0'; h_s) = \omega_j^{(1)}(\mathbf{d}_0, \mathbf{d}_0'; h_s) / \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \omega_j^{(1)}(\mathbf{d}_0, \mathbf{d}_m; h_s),$$

$$\Sigma_j^{(1)}(\mathbf{d}_0; h_s) = \mathbf{e}_{j,p}^T \Omega_X^{-1} \mathbf{e}_{j,p} \sum_{\mathbf{d}_m, \mathbf{d}_m' \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \tilde{\omega}_j^{(1)}(\mathbf{d}_0, \mathbf{d}_m; h_s) \tilde{\omega}_j^{(1)}(\mathbf{d}_0, \mathbf{d}_m'; h_s) \Sigma_y(\mathbf{d}_m, \mathbf{d}_m').$$

Theorem 4. *Suppose assumptions (C1)-(C9) and (C11) in Section 6 hold for piecewise continuous $\{\beta_{j*}(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$. For all $0 \le s \le S$, we have the following results:*

(i) $\sup_{\mathbf{d}_0 \in \mathcal{D}_0} |\tilde{\beta}_{j*}(\mathbf{d}_0; h_s) - \beta_{j*}(\mathbf{d}_0)| = O_p(h_s)$*;*

(ii) $\hat{\beta}_j(\mathbf{d}_0; h_s) - \tilde{\beta}_{j*}(\mathbf{d}_0; h_s) = \sum_{\mathbf{d}_m \in B(\mathbf{d}_0, h_s) \cap \mathcal{D}_0} \tilde{\omega}_j^{(1)}(\mathbf{d}_0, \mathbf{d}_m; h_s) \hat{\Delta}_j(\mathbf{d}_m)[1 + o_p(1)]$*;*

(iii) $\sup_{\mathbf{d}_0 \in \mathcal{D}_0} |\hat{\Sigma}(\sqrt{n}\tilde{\beta}_{j*}(\mathbf{d}_0; h_s)) - \Sigma_j^{(1)}(\mathbf{d}_0; h_s)| = o_p(1)$.

(iv) $\sqrt{n}[\hat{\beta}_j(\mathbf{d}_0; h_s) - \tilde{\beta}_{j*}(\mathbf{d}_0; h_s)]$ *converges in distribution to a normal distribution with mean zero and variance* $\Sigma_j^{(1)}(\mathbf{d}_0; h_s)$ *as* $n \to \infty$.

Remark 4. Theorem 4 characterizes several key features of MASS for a piecewise continuous function $\beta_{j*}(\mathbf{d})$. These results differ significantly from those for the piecewise constant case, but under weaker assumptions. For instance, Theorem 4 (i) quantifies the bias of the pseudo true value $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s)$ relative to the true value $\beta_{j*}(\mathbf{d}_0)$ across all $\mathbf{d}_0 \in \mathcal{D}_0$ for a fixed s. Even for voxels inside the smooth areas of $\beta_{j*}(\mathbf{d})$, the bias $O_p(h_s)$ is still much higher than the standard bias at the rate of h_s^2 due to the presence of $K_{st}(D_{\beta_j}(\mathbf{d}_0, \mathbf{d}_0'; h_{s-1})/C_n)$ (Fan and Gijbels, 1996; Wand and Jones, 1995). If we set $K_{st}(u) = \mathbf{1}(u \in [0,1])$ and $\beta_{j*}(\mathbf{d})$ is twice differentiable, then the bias of $\tilde{\beta}_{j*}(\mathbf{d}_0; h_s)$ relative to $\beta_{j*}(\mathbf{d}_0)$ may be reduced to $O_p(h_s^2)$. Theorem 4 (iv) ensures that

$\sqrt{n}[\hat{\beta}_j(\mathbf{d}_0; h_s) - \tilde{\beta}_{j*}(\mathbf{d}_0; h_s)]$ is asymptotically normally distributed. Moreover, as shown in the supplementary document, $\Sigma_j^{(1)}(\mathbf{d}_0; h_s)$ is smaller than the asymptotic variance of the raw estimate $\hat{\beta}_j(\mathbf{d}_0)$, and thus MASS can increase statistical power in testing $H_0(\mathbf{d}_0)$ even for the piecewise continuous case.

3 Simulation Studies

In this section, we conducted a set of Monte Carlo simulations to compare MASS with voxel-wise methods from three different aspects. Firstly, we examine the finite sample performance of $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_s)$ at different signal-to-noise ratios. Secondly, we examine the accuracy of the estimated eigenfunctions of $\Sigma_\eta(\mathbf{d}, \mathbf{d}')$. Thirdly, we assess both Type I and II error rates of the Wald test statistic. For the sake of space, we only present some selected results below and put additional simulation results in the supplementary document.

We simulated data at all 32,768 voxels on the $64 \times 64 \times 8$ phantom image for $n = 60$ (or 80) subjects. At each $\mathbf{d}_0 = (d_{0,1}, d_{0,2}, d_{0,3})^T$ in $\mathcal{D}_0$, $Y_i(\mathbf{d}_0)$ was simulated according to

$$y_i(\mathbf{d}_0) = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{d}_0) + \eta_i(\mathbf{d}_0) + \epsilon_i(\mathbf{d}_0) \quad \text{for} \quad i = 1, \ldots, n, \tag{25}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^T$, $\boldsymbol{\beta}(\mathbf{d}_0) = (\beta_1(\mathbf{d}_0), \beta_2(\mathbf{d}_0), \beta_3(\mathbf{d}_0))^T$, and $\epsilon(\mathbf{d}_0) \sim N(0, 1)$ or $\chi(3)^2 - 3$, in which $\chi^2(3) - 3$ is a very skewed distribution. Furthermore, we set $\eta_i(\mathbf{d}_0) = \sum_{l=1}^3 \xi_{il}\psi_l(\mathbf{d}_0)$, where ξ_{il} are independently generated according to $\xi_{i1} \sim N(0, 0.6)$, $\xi_{i2} \sim N(0, 0.3)$, and $\xi_{i3} \sim N(0, 0.1)$, $\psi_1(\mathbf{d}_0) = 0.5\sin(2\pi d_{0,1}/64)$, $\psi_2(\mathbf{d}_0) = 0.5\cos(2\pi d_{0,2}/64)$, and $\psi_3(\mathbf{d}_0) = \sqrt{1/2.625}(9/8 - d_{0,3}/4)$. The first eigenfunction $\psi_1(\mathbf{d}_0)$ changes only along $d_{0,1}$ direction, while it keeps constant in the other two directions. The other two eigenfunctions, $\psi_2(\mathbf{d}_0)$ and $\psi_3(\mathbf{d}_0)$, were chosen in a similar way (Figure 3). We set $x_{i1} = 1$ and generated x_{i2} independently from a Bernoulli distribution with success rate 0.5 and x_{i3} independently from the uniform distribution on $[1, 2]$. The covariates x_{i2} and x_{i3} were chosen to represent group identity and scaled age, respectively.

We chose different pattens for different $\beta_j(\mathbf{d})$ images in order to examine the finite

sample performance of our estimation method under different scenarios. We set all the 8 slices along the coronal axis to be identical for each of $\beta_j(\mathbf{d})$ images. As shown in Figure 4, each slice of the three different $\beta_j(\mathbf{d})$ images has four different blocks and 5 different regions of interest (ROIs) with varying patterns and shape. The true values of $\beta_j(\mathbf{d})$ were varied from 0 to 0.8, respectively, and were displayed for all ROIs with navy blue, blue, green, orange and brown colors representing $0, 0.2, 0.4, 0.6$, and 0.8, respectively.

We fitted the SVCM model (1) with the same set of covariates to a simulated data set, and then applied the three-stage estimation procedure described in Section 2.2 to calculate adaptive parameter estimates across all pixels at 11 different scales. In MASS, we set $h_s = 1.1^s$ for $s = 0, \ldots, S = 10$. Figure 4 shows some selected slices of $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_s)$ at $s = 0$ (middle panels) and $s = 10$ (lower panels). Inspecting Figure 4 reveals that all $\hat{\beta}_j(\mathbf{d}_0; h_{10})$ outperform their corresponding $\hat{\beta}_j(\mathbf{d}_0)$ in terms of variance and detected ROI patterns. Following the method described in Section 2.2, we estimated $\eta_i(\mathbf{d})$ based on the residuals $y_i(\mathbf{d}_0) - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\mathbf{d}_0)$ by using the local linear smoothing method and then calculate $\hat{\eta}_i(\mathbf{d})$. Figure 3 shows some selected slices of the first three estimated eigenfunctions. Inspecting Figure 3 reveals that $\hat{\eta}_i(\mathbf{d})$ are relatively close to the true eigenfunctions and can capture the main feature in the true eigenfunctions, which vary in one direction and are constant in the other two directions. However, we do observe some minor block effects, which may be caused by using the block smoothing method to estimate $\eta_i(\mathbf{d})$.

Furthermore, for $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_s)$, we calculated the bias, the empirical standard error (RMS), the mean of the estimated standard errors (SD), and the ratio of RMS over SD (RE) at each voxel of the five ROIs based on the results obtained from the 200 simulated data sets. For the sake of space, we only presented some selected results based on $\hat{\beta}_3(\mathbf{d}_0)$ and $\hat{\beta}_3(\mathbf{d}_0; h_{10})$ obtained from $N(0, 1)$ distributed data with $n = 60$ in Table 1. The biases are slightly increased from h_0 to h_{10} (Table 1), whereas RMS and SD at h_5 and h_{10} are much smaller than those at h_0 (Table 1). In addition, the RMS and its corresponding SD are relatively close to each other at all scales for both the normal and Chi-square distributed data (Table 1). Moreover, SDs in these voxels of ROIs with

$\beta_3(\mathbf{d}_0) > 0$ are larger than SDs in those voxels of ROI with $\beta_3(\mathbf{d}_0) = 0$, since the interior of ROI with $\beta_3(\mathbf{d}_0) = 0$ contains more pixels (Figure 4 (c)). Moreover, the SDs at steps h_0 and h_{10} show clear spatial patterns caused by spatial correlations. The RMSs also show some evidence of spatial patterns. The biases, SDs, and RMSs of $\beta_3(\mathbf{d}_0)$ are smaller in the normal distributed data than in the chi-square distributed data (Table 1), because the signal-to-noise ratios (SNRs) in the normal distributed data are bigger than those SNRs in the chi-square distributed data. Increasing sample size and signal-to-noise ratio decreases the bias, RMS and SD of parameter estimates (Table 1).

To assess both Type I and II error rates at the voxel level, we tested the hypotheses $H_0(\mathbf{d}_0) : \beta_j(\mathbf{d}_0) = 0$ versus $H_1(\mathbf{d}_0) : \beta_j(\mathbf{d}_0) \neq 0$ for $j = 1, 2, 3$ across all $\mathbf{d}_0 \in \mathcal{D}_0$. We applied the same MASS procedure at scales h_0 and h_{10}. The $-\log_{10}(p)$ values on some selected slices are shown in the supplementary document. The 200 replications were used to calculate the estimates (ES) and standard errors (SE) of rejection rates at $\alpha = 5\%$ significance level. Due to space limit, we only report the results of testing $\beta_2(\mathbf{d}_0) = 0$. The other two tests have similar results and are omitted here. For $W_\beta(\mathbf{d}_0; h)$, the Type I rejection rates in ROI with $\beta_2(\mathbf{d}_0) = 0$ are relatively accurate for all scenarios, while the statistical power for rejecting the null hypothesis in ROIs with $\beta_2(\mathbf{d}_0) \neq 0$ significantly increases with radius h_s and signal-to-noise ratio (Table 2). As expected, increasing n improves the statistical power for detecting $\beta_2(\mathbf{d}_0) \neq 0$.

4 Real Data Analysis

We applied SVCM to the Attention Deficit Hyperactivity Disorder (ADHD) data from the New York University (NYU) site as a part of the ADHD-200 Sample Initiative (http://fcon_1000.projects.nitrc.org/indi/adhd200/). ADHD-200 Global Competition is a grassroots initiative event to accelerate the scientific community's understanding of the neural basis of ADHD through the implementation of open data-sharing and discovery-based science. Attention deficit hyperactivity disorder (ADHD) is one of the most common childhood disorders and can continue through adolescence and adulthood (Polanczyk

et al., 2007). Symptoms include difficulty staying focused and paying attention, difficulty controlling behavior, and hyperactivity (over-activity). It affects about 3 to 5 percent of children globally and diagnosed in about 2 to 16 percent of school aged children (Polanczyk et al., 2007). ADHD has three subtypes, namely, predominantly hyperactive-impulsive type, predominantly inattentive type, and combined type.

The NYU data set consists of 174 subjects (99 Normal Controls (NC) and 75 ADHD subjects with combined hyperactive-impulsive). Among them, there are 112 males whose mean age is 11.4 years with standard deviation 7.4 years and 62 females whose mean age is 11.9 years with standard deviation 10 years. Resting-state functional MRIs and T1-weighted MRIs were acquired for each subject. We only use the T1-weighted MRIs here. We processed the T1-weighted MRIs by using a standard image processing pipeline detailed in the supplementary document. Such pipeline consists of AC (anterior commissure) and -PC (posterior commissure) correction, bias field correction, skull-stripping, intensity inhomogeneity correction, cerebellum removal, segmentation, and nonlinear registration. We segmented each brain into three different tissues including grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). We used the RAVENS maps to quantify the local volumetric group differences for the whole brain and each of the segmented tissue type (GM, WM, and CSF) respectively, using the deformation field that we obtained during registration (Davatzikos et al., 2001). RAVENS methodology is based on a volume-preserving spatial transformation, which ensures that no volumetric information is lost during the process of spatial normalization, since this process changes an individuals brain morphology to conform it to the morphology of the Jacob template (Kabani et al., 1998).

We fitted model (1) to the RAVEN images calculated from the NYU data set. Specifically, we set $\boldsymbol{\beta}(\mathbf{d}_0) = (\beta_1(\mathbf{d}_0), \ldots, \beta_8(\mathbf{d}_0))^T$ and $\mathbf{x}_i = (1, \mathrm{G}_i, \mathrm{A}_i, \mathrm{D}_i, \mathrm{WBV}_i, \mathrm{A}_i \times \mathrm{D}_i, \mathrm{G}_i \times \mathrm{D}_i, \mathrm{A}_i \times \mathrm{G}_i)^T$, where G_i, A_i, D_i, and WBV_i, respectively, represent gender, age, diagnosis (1 for NC and 0 for ADHD), and whole brain volume. We applied the three-stage estimation procedure described in Section 2.2. In MASS, we set $h_s = 1.1^s$ for $s =$

$1, \ldots, 10$. We are interested in assessing the age and diagnosis interaction and the gender and diagnosis interaction. Specifically, we tested $H_0(\mathbf{d}_0): \beta_6(\mathbf{d}_0) = 0$ against $H_1(\mathbf{d}_0): \beta_6(\mathbf{d}_0) \neq 0$ for the age×diagnosis interaction across all voxels. Moreover, we also tested $H_0(\mathbf{d}_0): \beta_7(\mathbf{d}_0) = 0$ against $H_1(\mathbf{d}_0): \beta_7(\mathbf{d}_0) \neq 0$ for the gender×diagnosis interaction, but we present the associated results in the supplementary document. Furthermore, as shown in the supplementary document, the largest estimated eigenvalue is much larger than all other estimated eigenvalues, which decrease very slowly to zero, and explains 22% of variation in data after accounting for $\mathbf{x}_i$. Inspecting Figure 5 reveals that the estimated eigenfunction corresponding to the largest estimated eigenvalue captures the dominant morphometric variation.

As s increases from 0 to 10, MASS shows an advantage in smoothing effective signals within relatively homogeneous ROIs, while preserving the edges of these ROIs (Fig. 6 (a)-(d)). Inspecting Figure 6 (c) and (d) reveals that it is much easier to identify significant ROIs in the $-\log_{10}(p)$ images at scale h_{10}, which are much smoother than those at scale h_0. To formally detect significant ROIs, we used a cluster-form of threshold of 5% with a minimum voxel clustering value of 50 voxels. We were able to detect 26 significant clusters across the brain. Then, we overlapped these clusters with the 96 predefined ROIs in the Jacob template and were able to detect several predefined ROIs for each cluster. As shown in the supplementary document, we were able to detect several major ROIs, such as the frontal lobes and the right parietal lobe. The anatomical disturbance in the frontal lobes and the right parietal lobe has been consistently revealed in the literature and may produce difficulties with inhibiting prepotent responses and decreased brain activity during inhibitory tasks in children with ADHD (Bush, 2011). These ROIs comprise the main components of the cingulo-frontal-parietal cognitive-attention network. These areas, along with striatum, premotor areas, thalamus and cerebellum have been identified as nodes within parallel networks of attention and cognition (Bush, 2011).

To evaluate the prediction accuracy of SVCM, we randomly selected one subject with ADHD from the NYU data set and predicted his/her RAVENS image by using

both model (1) and a standard linear model with normal noise. In both models, we used the same set of covariates, but different covariance structures. Specifically, in the standard linear model, an independent correlation structure was used and the least squares estimates of $\boldsymbol{\beta}(\mathbf{d}_0)$ were calculated. For SVCM, the functional principal component analysis model was used and $\hat{\boldsymbol{\beta}}(\mathbf{d}_0; h_{10})$ were calculated. After fitting both models to all subjects except the selected one, we used the fitted models to predict the RAVEN image of the selected subject and then calculated the prediction error based on the difference between the true and predicted RAVEN images. We repeated the prediction procedure 50 times and calculated the mean and standard deviation images of these prediction error images (Figure 7). Inspecting Figure 7 reveals the advantage and accuracy of model (1) over the standard linear model for the ADHD data.

5 Discussion

This article studies the idea of using SVCM for the spatial and adaptive analysis of neuroimaging data with jump discontinuities, while explicitly modeling spatial dependence in neuroimaging data. We have developed a three-stage estimation procedure to carry out statistical inference under SVCM. MASS integrates three methods including propagation-separation, functional principal component analysis, and jumping surface model for neuroimaging data from multiple subjects. We have developed a fast and accurate estimation method for independently updating each of effect images, while consistently estimating their standard deviation images. Moreover, we have derived the asymptotic properties of the estimated eigenvalues and eigenfunctions and the parameter estimates.

Many issues still merit further research. The basic setup of SVCM can be extended to more complex data structures (e.g., longitudinal, twin and family) and other parametric and semiparametric models. For instance, we may develop a spatial varying coefficient mixed effects model for longitudinal neuroimaging data. It is also feasible to include nonparametric components in SVCM. More research is needed for weakening

regularity assumptions and for developing adaptive-neighborhood methods to determine multiscale neighborhoods that adapt to the pattern of imaging data at each voxel. It is also interesting to examine the efficiency of our adaptive estimators obtained from MASS for different kernel functions and coefficient functions. An important issue is that SVCM and other voxel-wise methods do not account for the errors caused by registration method. We may need to explicitly model the measurement errors caused by the registration method, and integrate them with smoothing method and SVCM into a unified framework.

6 Technical Conditions

6.1 Assumptions

Throughout the paper, the following assumptions are needed to facilitate the technical details, although they may not be the weakest conditions. We do not distinguish the differentiation and continuation at the boundary points from those in the interior of $\mathcal{D}$.

Assumption C1. The number of parameters p is finite. Both N_D and n increase to infinity such that $\lim_{n\to\infty} C_n/n = \lim_{n\to\infty} C_n^{-1}\log(N_D) = \lim_{n\to\infty} C_n^{-1} = 0$.

Assumption C2. $\epsilon_i(\mathbf{d})$ are identical and independent copies of $\mathrm{SP}(0, \Sigma_\epsilon)$ and $\epsilon_i(\mathbf{d})$ and $\epsilon_i(\mathbf{d}')$ are independent for $\mathbf{d} \neq \mathbf{d}' \in \mathcal{D}$. Moreover, $\epsilon_i(\mathbf{d})$ are, uniformly in d, sub-Gaussian such that $K_\epsilon^2[E\exp(|\epsilon_i(\mathbf{d})|^2/K_\epsilon) - 1] \leq C_\epsilon$ for all $\mathbf{d} \in \mathcal{D}$ and some positive constants K_ϵ and C_ϵ.

Assumption C3. The covariate vectors $\mathbf{x}_i$s are independently and identically distributed with $E\mathbf{x}_i = \mu_x$ and $||\mathbf{x}_i||_\infty < \infty$. Moreover, $E(\mathbf{x}_i^{\otimes 2}) = \Omega_X$ is invertible. The $\mathbf{x}_i$, $\epsilon_i(\mathbf{d})$, and $\eta_i(\mathbf{d})$ are mutually independent of each other.

Assumption C4. Each component of $\{\eta(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$, $\{\eta(\mathbf{d})\eta(\mathbf{d}')^T : (\mathbf{d}, \mathbf{d}') \in \mathcal{D}^2\}$ and $\{\mathbf{x}\eta^T(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$ are Donsker classes. Moreover, $\min_{\mathbf{d}\in\mathcal{D}} \Sigma_\eta(\mathbf{d}, \mathbf{d}) > 0$ and $E[\sup_{\mathbf{d}\in\mathcal{D}} ||\eta(\mathbf{d})||_2^{2r_1}] < \infty$ for some $r_1 \in (2, \infty)$, where $||\cdot||_2$ is the Euclidean norm.

All components of $\Sigma_\eta(\mathbf{d}, \mathbf{d}')$ have continuous second-order partial derivatives with respect to $(\mathbf{d}, \mathbf{d}') \in \mathcal{D}^2$.

Assumption C5. The grid points $\mathcal{D}_0 = \{\mathbf{d}_m, m = 1, \ldots, N_D\}$ are independently and identically distributed with density function $\pi(\mathbf{d})$, which has the bounded support $\mathcal{D}$. Moreover, $\pi(\mathbf{d}) > 0$ for all $\mathbf{d} \in \mathcal{D}$ and $\pi(\mathbf{d})$ has continuous second-order derivative.

Assumption C6. The kernel functions $K_{loc}(t)$ and $K_{st}(t)$ are Lipschitz continuous and symmetric density functions, while $K_{loc}(t)$ has a compact support $[-1, 1]$. Moreover, they are continuously decreasing functions of $t \geq 0$ such that $K_{st}(0) = K_{loc}(0) > 0$ and $\lim_{t\to\infty} K_{st}(t) = 0$.

Assumption C7. h converges to zero such that

$$h \geq c(\log N_D/N_D)^{1-2/q_1} \quad \text{and} \quad h^{-12}(\log n/n)^{1-1/q_2} = o(1),$$

where $c > 0$ is a fixed constant and $\min(q_1, q_2) > 2$.

Assumption C8. There is a positive integer $E < \infty$ such that $\lambda_1 > \ldots > \lambda_E \geq 0$.

Assumption C9. For each j, the three assumptions of the jumping surface model hold, each $D^o_{j,l}$ is path-connected, and $\beta_{j*}(\mathbf{d})$ is a Lipschitz function of $\mathbf{d}$ with a common Lipschitz constant $K_j > 0$ in each $\mathcal{D}^o_{j,l}$ such that $|\beta_{j*}(\mathbf{d}) - \beta_{j*}(\mathbf{d}')| \leq K_j||\mathbf{d} - \mathbf{d}'||_2$ for any $\mathbf{d}, \mathbf{d}' \in \mathcal{D}^o_{j,l}$. Moreover, $\sup_{\mathbf{d}\in\mathcal{D}} |\beta_{j*}(\mathbf{d})| < \infty$, and $\max(K_j, L_j) < \infty$.

Assumption C10. For piecewise constant $\beta_{j*}(\mathbf{d})$, $o(\boldsymbol{u}^{(j)}(h_s)) = \sqrt{\log(1 + N_D)/n}$ and $N_D h_s^3 K_{st}(C_n^{-1} n \boldsymbol{u}^{(j)}(h_s)^2/(3S_y)) = o(\sqrt{\log(1 + N_D)/n})$ holds uniformly for $h_0 = 0 < \cdots < h_S$, where $S_y = \max_{\mathbf{d}_0\in\mathcal{D}_0} \Sigma_y(\mathbf{d}_0, \mathbf{d}_0)$ and $\boldsymbol{u}^{(j)}(h_s)$ is the smallest absolute value of all possible jumps at scale h_s and given by

$$\boldsymbol{u}^{(j)}(h_s) = \min\{|\beta_{j*}(\mathbf{d}_0) - \beta_{j*}(\mathbf{d}_0')| : (\mathbf{d}_0, \mathbf{d}_0') \in \mathcal{D}_0^2, \beta_{j*}(\mathbf{d}_0) \neq \beta_{j*}(\mathbf{d}_0'), \mathbf{d}_0' \in B(\mathbf{d}_0, h_s)\}.$$

Assumption C11. For piecewise continuous $\beta_{j*}(\mathbf{d})$, $\cup_{\mathbf{d}\in\mathcal{D}_0}[P_j(\mathbf{d}_0, h_S)^c \cap I_j(\mathbf{d}_0, \delta_L, \delta_U)]$ is an empty set and $h_0 = 0 < h_1 < \cdots < h_S$ is a sequence of bandwidths such that $\delta_L = O(\sqrt{\log(1+N_D)/n}) = o(1)$, $\delta_U = \sqrt{C_n/n}M_n = o(1)$, in which $\lim_{n\to\infty} M_n = \infty$, $h_S = O(\sqrt{\log(1+N_D)/n})$ and $N_D h_S^3 K_{st}(M_n^2/(3S_y)) = o(\sqrt{\log(1+N_D)/n})$.

REMARK 5. Assumption (C2) is needed to invoke Hoeffding inequality (Buhlmann and van de Geer, 2011; van der Vaar and Wellner, 1996) in order to establish the uniform bound for $\hat{\beta}(\mathbf{d}_0; h_s)$. In practice, since most neuroimaging data are often bounded, the sub-Gaussian assumption is reasonable. The bound assumption on $||\mathbf{x}||_\infty$ in Assumption (C3) is not essential and can be removed if we put a restriction on the tail of the distribution $\mathbf{x}$. Moreover, with some additional efforts, all results are valid even for the case with fixed design predictors. Assumption (C4) avoids smoothness conditions on the sample path $\eta(\mathbf{d})$, which are commonly assumed in the literature (Hall et al., 2006). The assumption on the moment of $\sup_{\mathbf{d}\in\mathcal{D}} ||\eta(\mathbf{d})||_2^{2r_2}$ is similar to the conditions used in (Li and Hsing, 2010). Assumption (C5) on the stochastic grid points is not essential and can be modified to accommodate the case for fixed grid points with some additional complexities.

REMARK 6. The bounded support restriction on $K_{loc}(\cdot)$ in Assumption (C6) can be weaken to a restriction on the tails of $K_{loc}(\cdot)$. Assumption (C9) requires smoothness and shape conditions on the image of $\beta_{j*}(\mathbf{d})$ for each j. For piecewise constant $\beta_{j*}(\mathbf{d})$, assumption (C10) requires conditions on the amount of changes at jumping points relative to n, N_D, and h_S. If $K_{st}(t)$ has a compact support, then $K_{st}(\boldsymbol{u}^{(j)2}/C) = 0$ for relatively large $\boldsymbol{u}^{(j)2}$. In this case, h_S can be very large. However, for piecewise continuous $\beta_{j*}(\mathbf{d})$, assumption (C11) requires the convergence rate of h_S and the amount of changes at jumping points.

References

Besag, J. E. (1986), "On the statistical analysis of dirty pictures (with discussion)," *Journal of the Royal Statistical Society, Ser. B.*, 48,, 259–302.

Buhlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, New York, N.Y.: Springer.

Bush, G. (2011), "Cingulate, frontal and parietal cortical dysfunction in attention-deficit/hyperactivity disorder," *Bio Psychiatry*, 69, 1160–1167.

Chan, T. F. and Shen, J. (2005), *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, Philadelphia: SIAM.

Chumbley, J., Worsley, K, J., Flandin, G., and Friston, K. J. (2009), "False discovery rate revisited: FDR and topological inference using Gaussian random fields," *Neuroimage*, 44, 62–70.

Cressie, N. and Wikle, C. (2011), *Statistics for Spatio-Temporal Data,*, Hoboken, NJ: Wiley.

Davatzikos, C., Genc, A., Xu, D., and Resnick, S. (2001), "Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy." *NeuroImage*, 14, 1361–1369.

Fan, J. (1993), "Local linear regression smoothers and their minimax efficiencies," *Ann. Statist.*, 21, 196–216.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J. and Zhang, J. (2002), "Two-step estimation of functional linear models with applications to longitudinal data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 303–322.

Fan, J. and Zhang, W. (1999), "Statistical estimation in varying coefficient models," *The Annals of Statistics*, 27, 1491–1518.

— (2008), "Statistical methods with varying coefficient models," *Stat. Interface*, 1, 179–195.

Friston, K. J. (2007), *Statistical Parametric Mapping: the Analysis of Functional Brain Images*, London: Academic Press.

Hall, P., Müller, H.-G., and Wang, J.-L. (2006), "Properties of principal component methods for functional and longitudinal data analysis," *Ann. Statist.*, 34, 1493–1517.

Kabani, N., MacDonald, D., Holmes, C., and Evans, A. (1998), "A 3D atlas of the human brain," *Neuroimage*, 7, S717.

Khodadadi, A. and Asgharian, M. (2008), "Change point problem and regression: an annotated bibliography," Tech. rep., McGill University, http://biostats.bepress.com/cobra/art44.

Lazar, N. A. (2008), *The Statistical Analysis of Functional MRI Data*, New York: Springer.

Li, S. Z. (2009), *Markov Random Field Modeling in Image Analysis*, New York, NY: Springer.

Li, Y. and Hsing, T. (2010), "Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data," *The Annals of Statistics*, 38, 3321–3351.

Li, Y., Zhu, H., Shen, D., Lin, W., Gilmore, J. H., and Ibrahim, J. G. (2011), "Multiscale adaptive regression models for neuroimaging data," *Journal of the Royal Statistical Society: Series B*, 73, 559–578.

Liu, S. (1999), "Matrix results on the Khatri-Rao and Tracy-Singh products," *Linear Algebra Appl.*, 289, 267–277.

Mori, S. (2002), "Principles, methods, and applications of diffusion tensor imaging," *In Toga AW, Mazziotta JC, editors. Brain Mapping: The Methods, 2nd Edition. Elsevier Science*, 379–397.

Polanczyk, G., de Lima, M., Horta, B., Biederman, J., and Rohde, L. (2007), "The worldwide prevalence of ADHD: a systematic review and metaregression analysis," *The American Journal of Psychiatry*, 164, 942–948.

Polzehl, J. and Spokoiny, V. G. (2000), "Adaptive weights smoothing with applications to image restoration," *J. R. Statist. Soc. B*, 62, 335–354.

— (2006), "Propagation-separation approach for local likelihood estimation," *Probab. Theory Relat. Fields*, 135, 335–362.

Polzehl, J., Voss, H. U., and Tabelow, K. (2010), "Structural adaptive segmentation for statistical parametric mapping," *NeuroImage*, 52, 515–523.

Qiu, P. (2005), *Image Processing and Jump Regression Analysis*, New York: John Wiley & Sons.

— (2007), "Jump surface estimation, edge detection, and image restoration," *Journal of American Statistical Association*, 102, 745–756.

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

Scott, D. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.

Spence, J., Carmack, P., Gunst, R., Schucany, W., Woodward, W., and Haley, R. (2007), "Accounting for spatial dependence in the analysis of SPECT brain imaging data." *Journal of the American Statistical Association*, 102, 464–473.

Tabelow, K., Polzehl, J., Spokoiny, V., and Voss, H. U. (2008a), "Diffusion tensor imaging: structural adaptive smoothing," *NeuroImage*, 39, 1763–1773.

Tabelow, K., Polzehl, J., Ulug, A. M., Dyke, J. P., Watts, R., Heier, L. A., and Voss, H. U. (2008b), "Accurate localization of brain activity in presurgical fMRI by structure adaptive smoothing," *IEEE Trans. Med. Imaging*, 27, 531–537.

Thompson, P. and Toga, A. (2002), "A framework for computational anatomy," *Computing and Visualization in Science*, 5, 13–34.

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

Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., and Lerch, J. (2004), "Unified univariate and multivariate random field theory," *NeuroImage*, 23, 189–195.

Wu, C. O., Chiang, C. T., and Hoover, D. R. (1998), "Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data." *J. Amer. Statist. Assoc.*, 93, 1388–1402.

Yue, Y., Loh, J. M., and Lindquist, M. A. (2010), "Adaptive spatial smoothing of fMRI images." *Statistics and its Interface*, 3, 3–14.

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. (2011), "Functional principal component model for high-dimensional brain imaging," *NeuroImage*, 58, 772–784.

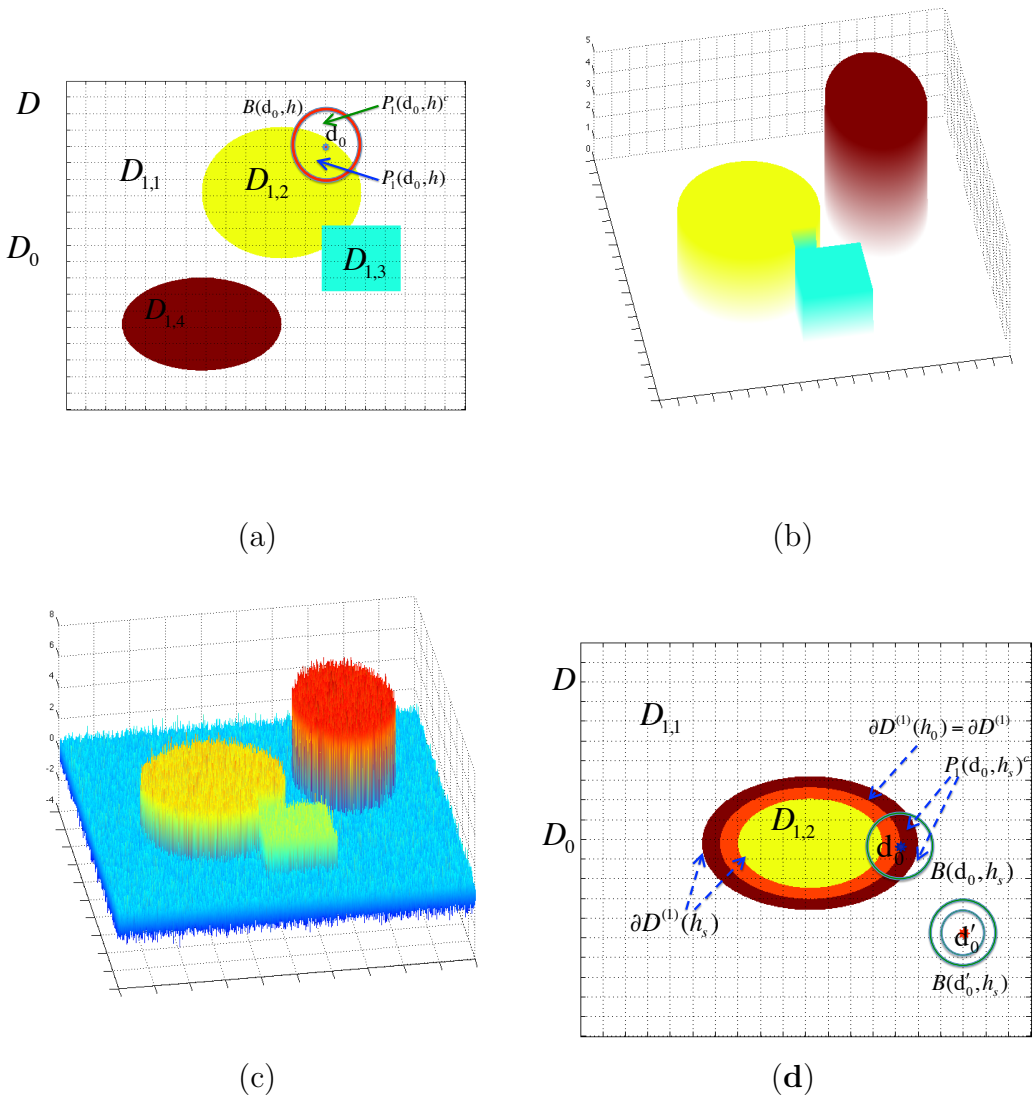

Figure 1: Illustration of a jumping surface model for $\beta_1(\mathbf{d})$ and boundary sets over a two-dimensional region D: (a) $\mathcal{D}$, $\mathcal{D}_0$, a disjoint partition of $\mathcal{D}$ as the union of four disjoint regions with white, yellow, blue green, and red representing $\mathcal{D}_{1,1}$, $\mathcal{D}_{1,2}$, $\mathcal{D}_{1,3}$, and $\mathcal{D}_{1,4}$, a representative voxel $\mathbf{d}_0 \in \mathcal{D}_0$, an open ball of $\mathbf{d}_0$, $B(\mathbf{d}_0, h)$, a maximal path-connected set $P_1(\mathbf{d}_0, h)$, and $P_1(\mathbf{d}_0, h)^c$; (b) three-dimensional shaded surface of true $\{\beta_1(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$ map; (c) three-dimensional shaded surface of estimated $\{\hat{\beta}_1(\mathbf{d}_0) : \mathbf{d}_0 \in \mathcal{D}_0\}$ map; and (d) $\mathcal{D}$, $\mathcal{D}_0$, a disjoint partition of $\mathcal{D} = \mathcal{D}_{1,1} \cup \mathcal{D}_{1,2}$, $\partial D^{(1)}(h_0) \subset \partial D^{(1)}(h_s)$, two representative voxels $\mathbf{d}_0$ and $\mathbf{d}'_0$ in $\mathcal{D}_0$, two open balls of $\mathbf{d}'_0 \in \mathcal{D}_{1,1}$, an open ball of $\mathbf{d}_0 \in \partial D^{(1)}(h_s) \cap \mathcal{D}_0$, $B(\mathbf{d}_0, h_s)$, and $P_1(\mathbf{d}_0, h_s)^c$.

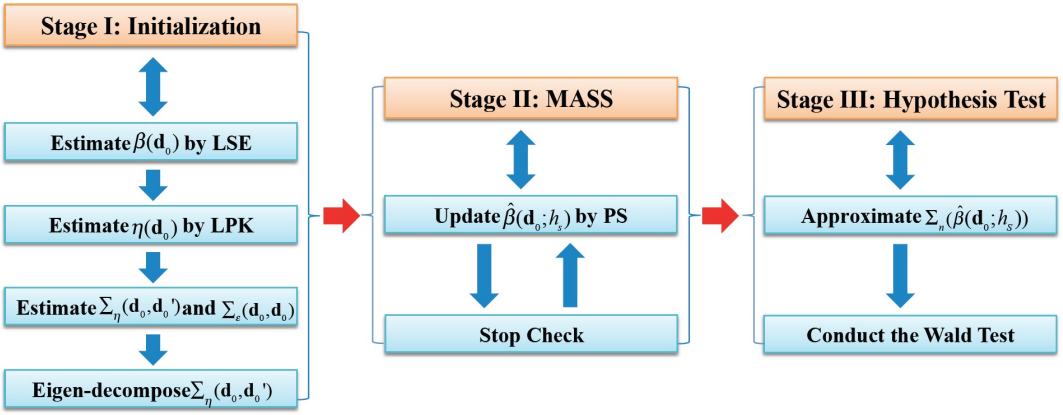

Figure 2: A schematic overview of the three stages of SVCM: Stage (I) is the initialization step, Stage (II) is the Multiscale Adaptive and Sequential Smoothing (MASS) method, and Stage (III) is the hypothesis test.

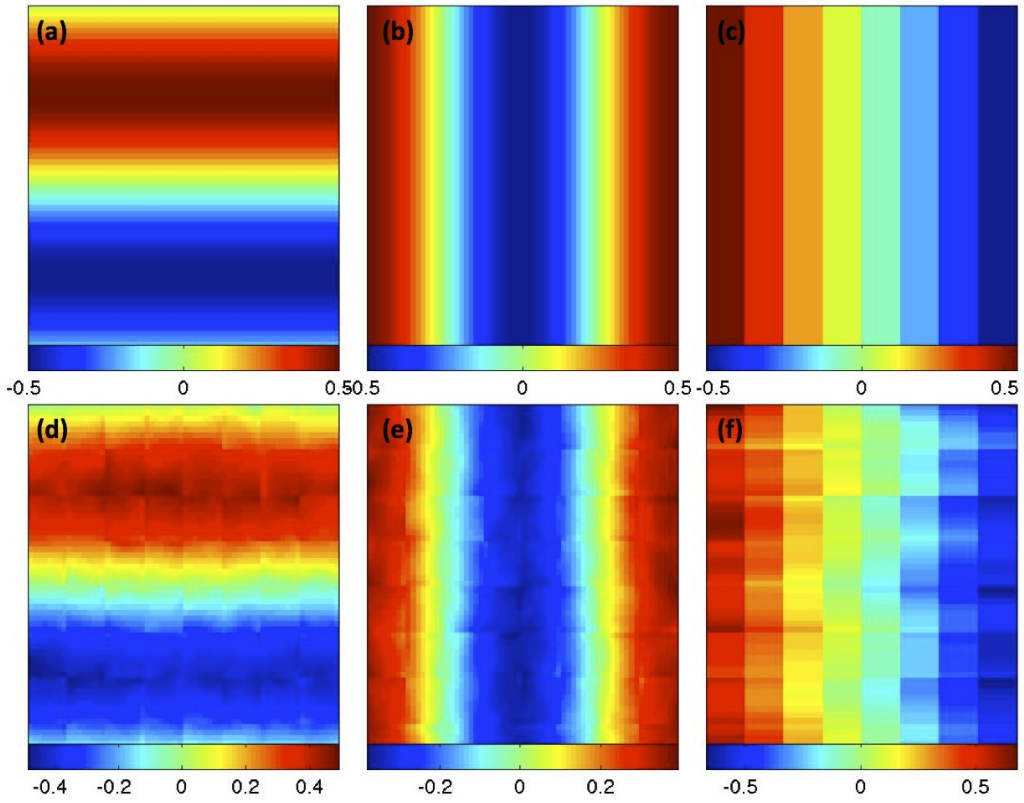

Figure 3: Simulation results: a selected slice of (a) true $\psi_1(\mathbf{d})$; (b) true $\psi_2(\mathbf{d})$; (c) true $\psi_3(\mathbf{d})$; (d) $\hat{\psi}_1(\mathbf{d})$; (e) $\hat{\psi}_2(\mathbf{d})$; and (f) $\hat{\psi}_3(\mathbf{d})$.

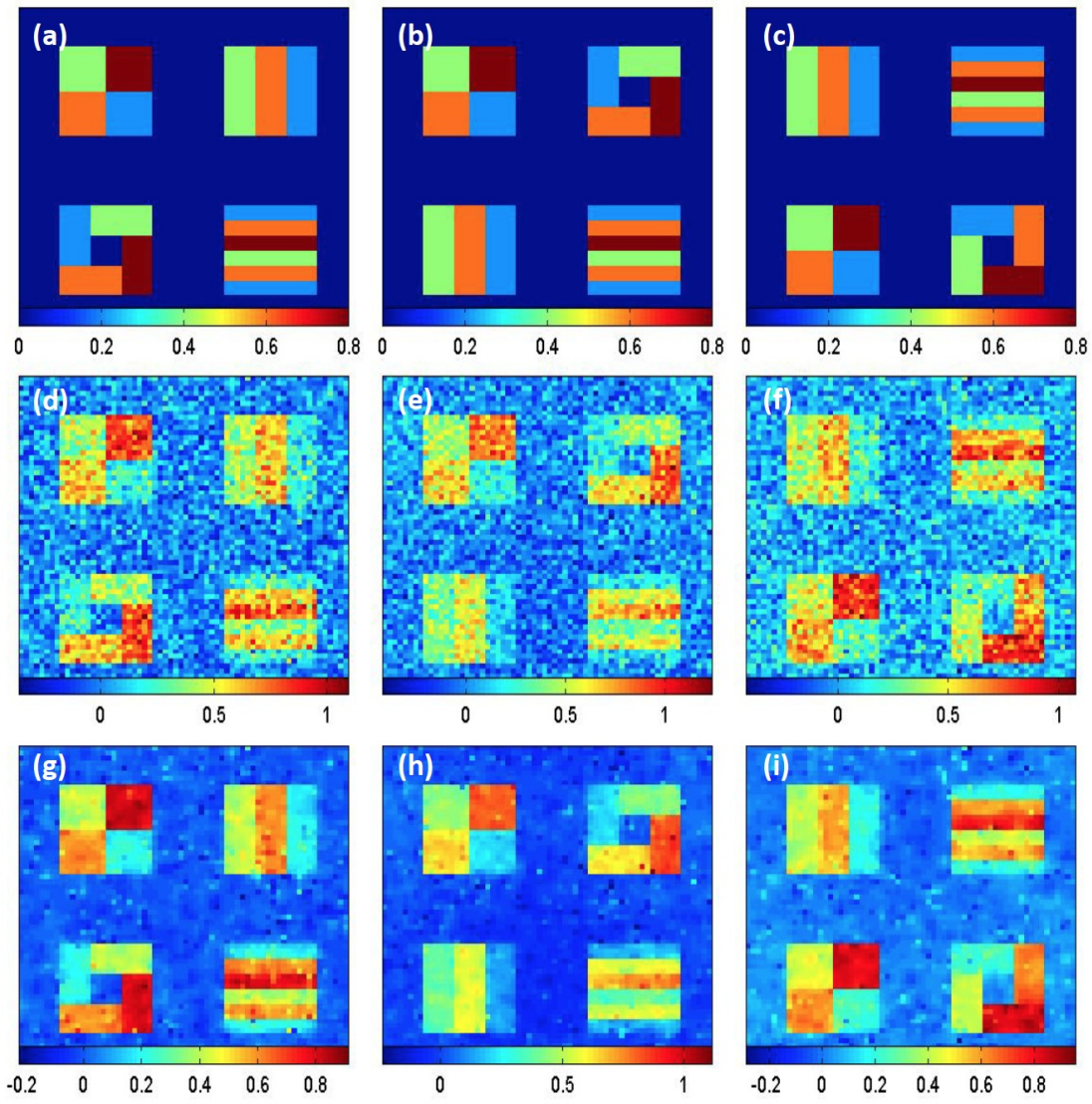

Figure 4: Simulation results: a selected slice of (a) true $\beta_1(\mathbf{d})$; (b) true $\beta_2(\mathbf{d})$; (c) true $\beta_3(\mathbf{d})$; (d) $\hat{\beta}_1(\mathbf{d}_0)$; (e) $\hat{\beta}_2(\mathbf{d}_0)$; (f) $\hat{\beta}_3(\mathbf{d}_0)$; (g) $\hat{\beta}_1(\mathbf{d}_0; h_{10})$; (h) $\hat{\beta}_2(\mathbf{d}_0; h_{10})$; and (i) $\hat{\beta}_3(\mathbf{d}_0; h_{10})$.

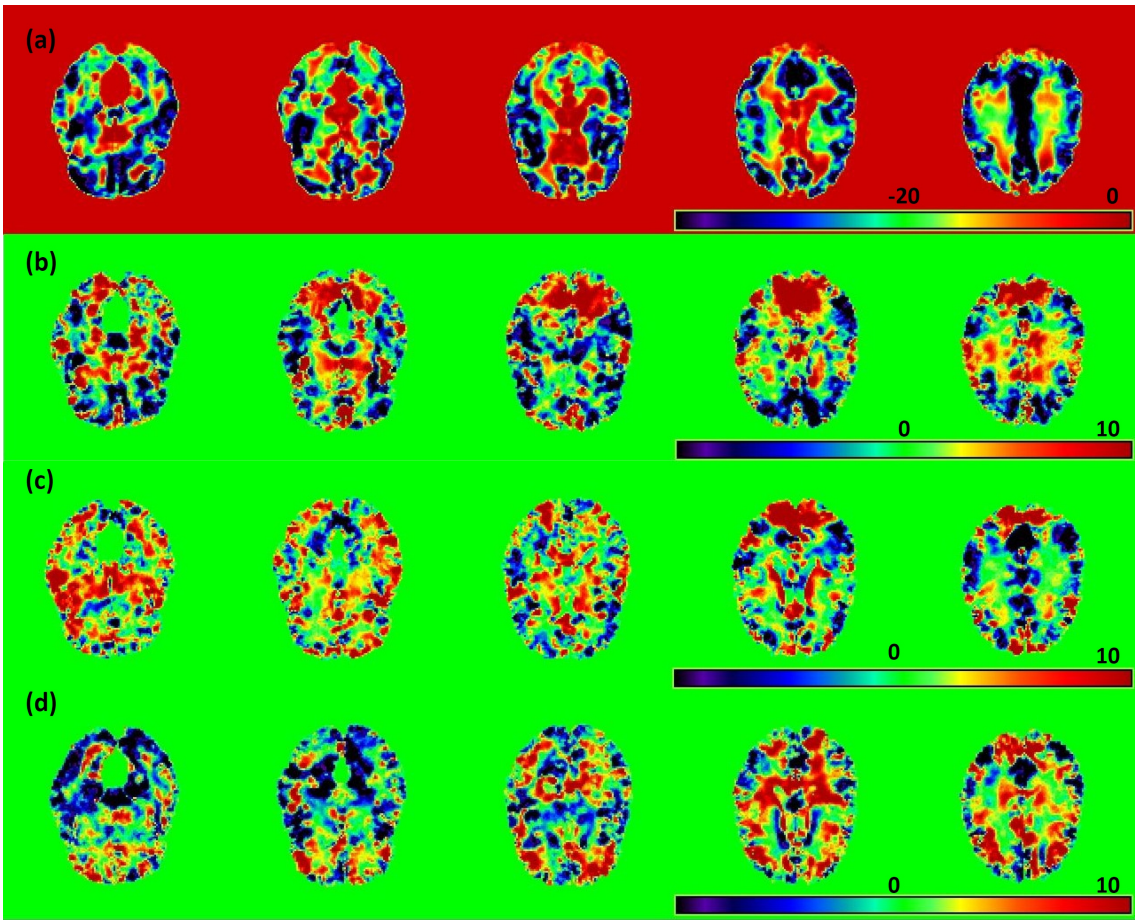

Figure 5: Results from the ADHD 200 data: five selected slices of the four estimated eigenfunctions corresponding to the first four largest eigenvalues of $\hat{\Sigma}_\eta(\cdot,\cdot)$: (a) $\hat{\psi}_1(\mathbf{d})$; (b) $\hat{\psi}_2(\mathbf{d})$; (c) $\hat{\psi}_3(\mathbf{d})$; and (d) $\hat{\psi}_4(\mathbf{d})$.

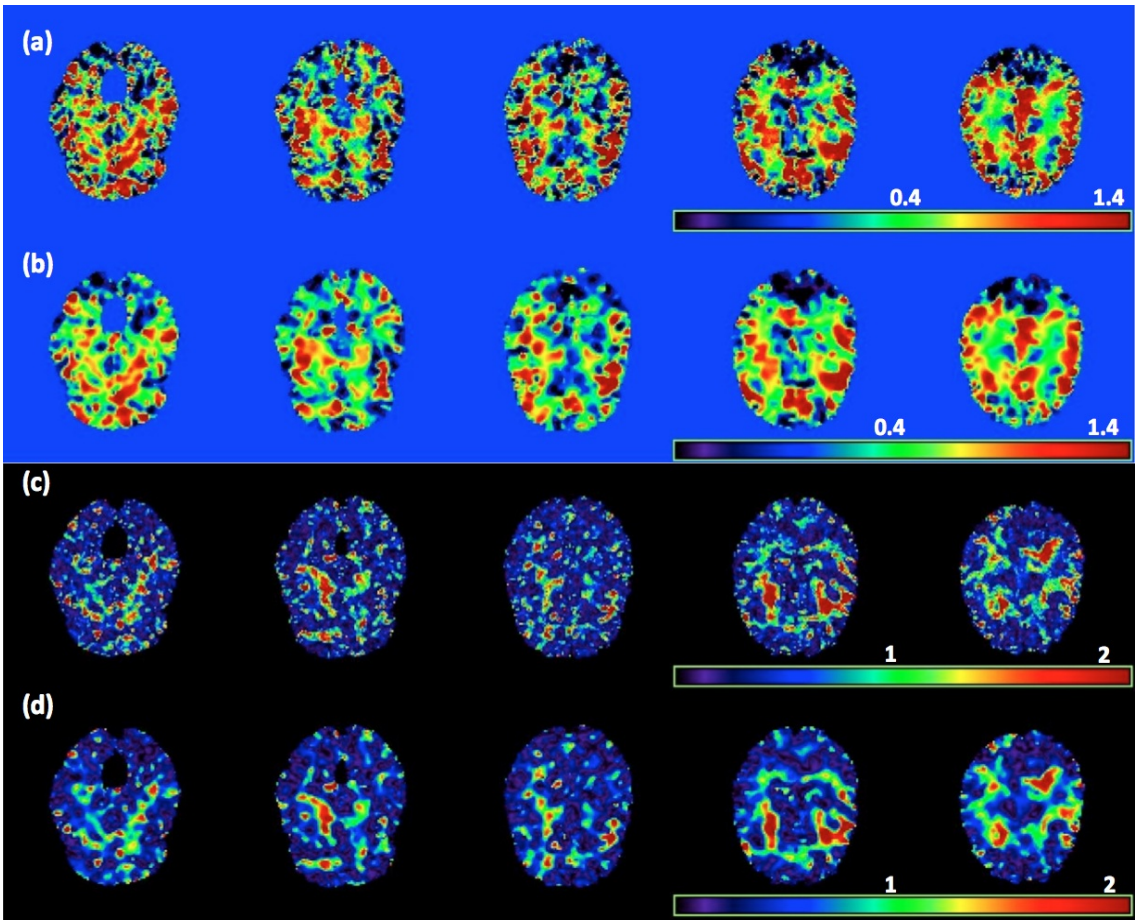

Figure 6: Results from the ADHD 200 data: five selected slices of (a) $\hat{\beta}_6(\mathbf{d}_0)$, (b) $\hat{\beta}_6(\mathbf{d}_0; h_{10})$, the $-\log_{10}(p)$ images for testing $H_0 : \beta_6(\mathbf{d}_0) = 0$ (c) at scale h_0 and (d) at scale h_{10}, where $\beta_6(\mathbf{d}_0)$ is the regression coefficient associated with the age×diagnostic interaction.

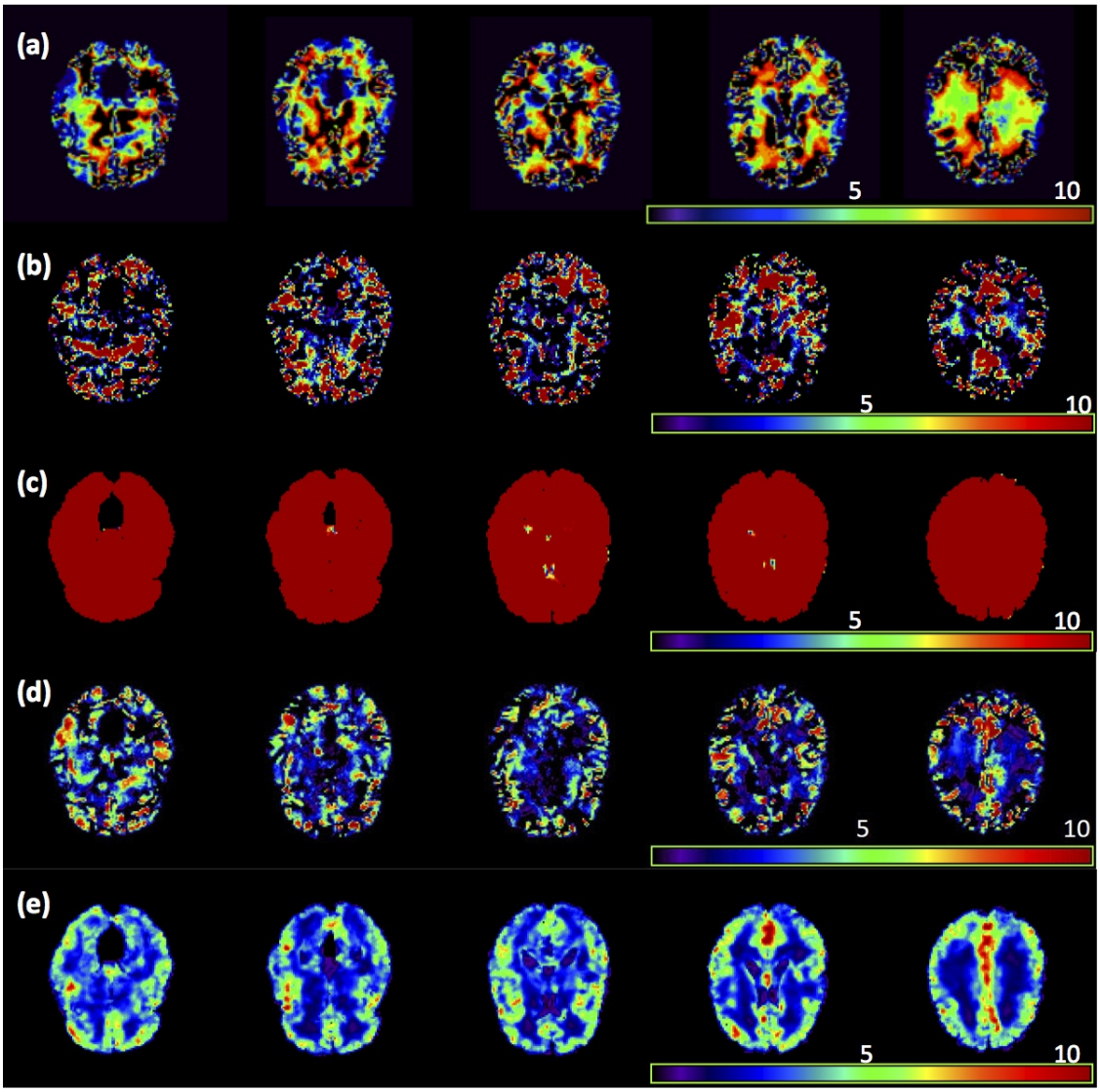

Figure 7: Results from the ADHD 200 data: The raw RAVENS image for a selected subject with ADHD (a), mean ((b) GLM and (d) SVCM) and standard error ((c) GLM and (e) SVCM) of the errors to predict the RAVENS image in (a), where GLM denotes general linear model.

Table 1: Simulation results: Average Bias ($\times 10^{-2}$), RMS, SD, and RE of $\beta_2(\mathbf{d}_0)$ parameters in the five ROIs at 3 different scales (h_0, h_5, h_{10}), $N(0,1)$ and $\chi(3)^2 - 3$ distributed noisy data, and 2 different sample sizes $(n = 60, 80)$. BIAS denotes the bias of the mean of estimates; RMS denotes the root-mean-square error; SD denotes the mean of the standard deviation estimates; RE denotes the ratio of RMS over SD. For each case, 200 simulated data sets were used.

		$\chi^2(3)-3$						$N(0,1)$					
		$n=60$			$n=80$			$n=60$			$n=80$		
$\beta_2(\mathbf{d}_0)$		h_0	h_5	h_{10}	h_0	h_5	h_{10}	h_0	h_5	h_{10}	h_0	h_5	h_{10}
0.0	BIAS	-0.03	0.36	0.61	0.00	0.34	0.56	-0.01	0.17	0.22	0.01	0.16	0.20
	RMS	0.18	0.13	0.13	0.15	0.10	0.10	0.14	0.07	0.07	0.12	0.06	0.06
	SD	0.18	0.13	0.12	0.15	0.11	0.11	0.14	0.07	0.07	0.12	0.06	0.06
	RE	1.03	1.00	1.04	1.00	0.94	0.98	0.99	0.94	1.03	1.00	0.95	1.04
0.2	BIAS	0.72	0.37	0.38	0.15	-0.35	-0.39	-0.04	-0.55	-0.66	0.10	-0.48	-0.61
	RMS	0.19	0.14	0.13	0.16	0.11	0.11	0.14	0.07	0.07	0.12	0.06	0.06
	SD	0.18	0.14	0.13	0.16	0.12	0.11	0.14	0.08	0.07	0.12	0.07	0.06
	RE	1.02	0.99	1.03	1.00	0.96	0.99	0.99	0.96	1.04	1.00	0.97	1.06
0.4	BIAS	-0.40	-0.55	-0.68	-0.10	-0.15	-0.24	0.04	0.12	0.13	-0.10	0.05	0.08
	RMS	0.19	0.14	0.14	0.16	0.12	0.12	0.14	0.07	0.07	0.12	0.07	0.07
	SD	0.18	0.14	0.13	0.16	0.12	0.12	0.14	0.08	0.07	0.12	0.07	0.06
	RE	1.02	1.00	1.03	1.00	0.96	1.00	0.99	0.96	1.04	1.00	0.97	1.06
0.6	BIAS	0.42	-1.14	-1.93	0.05	-1.20	-1.89	0.03	-0.55	-0.69	-0.01	-0.43	-0.54
	RMS	0.18	0.13	0.13	0.15	0.11	0.11	0.14	0.07	0.07	0.12	0.06	0.06
	SD	0.18	0.13	0.13	0.15	0.11	0.11	0.14	0.08	0.07	0.12	0.07	0.06
	RE	1.02	1.00	1.04	1.00	0.95	0.99	0.99	0.97	1.05	1.00	0.97	1.05
0.8	BIAS	-1.04	-2.95	-4.09	-0.13	-1.71	-2.70	-0.11	-0.82	-1.03	-0.03	-0.59	-0.77
	RMS	0.19	0.15	0.15	0.16	0.12	0.12	0.14	0.08	0.07	0.12	0.07	0.07
	SD	0.19	0.15	0.14	0.16	0.13	0.12	0.14	0.08	0.07	0.12	0.07	0.06
	RE	1.02	1.00	1.03	1.00	0.96	0.99	0.99	0.94	1.01	1.00	0.95	1.02

Table 2: Simulation Study for $W_\beta(\mathbf{d}_0; h)$: estimates (ES) and standard errors (SE) of rejection rates for pixels inside the five ROIs were reported at 2 different scales (h_0, h_{10}), $N(0,1)$ and $\chi^2(3) - 3$ distributed data, and 2 different sample sizes $(n = 60, 80)$ at $\alpha = 5\%$. For each case, 200 simulated data sets were used.

		$\chi^2(3)-3$				$N(0,1)$			
		$n=60$		$n=80$		$n=60$		$n=80$	
$\beta_2(\mathbf{d}_0)$	s	ES	SE	ES	SE	ES	SE	ES	SE
0.0	h_0	0.056	0.016	0.049	0.015	0.048	0.015	0.050	0.016
	h_{10}	0.055	0.016	0.042	0.015	0.036	0.016	0.040	0.019
0.2	h_0	0.210	0.043	0.245	0.039	0.282	0.033	0.370	0.035
	h_{10}	0.358	0.126	0.413	0.139	0.777	0.107	0.870	0.081
0.4	h_0	0.556	0.072	0.692	0.054	0.794	0.030	0.895	0.024
	h_{10}	0.792	0.129	0.894	0.078	0.994	0.006	0.998	0.003
0.6	h_0	0.907	0.040	0.966	0.022	0.988	0.008	0.998	0.003
	h_{10}	0.986	0.023	0.997	0.009	1.000	0.001	1.000	0.000
0.8	h_0	0.978	0.016	0.997	0.004	1.000	0.001	1.000	0.000
	h_{10}	0.997	0.006	1.000	0.001	1.000	0.000	1.000	0.000