Published in final edited form as:
*J R Stat Soc Series B Stat Methodol*. 2011 September ; 73(4): 559–578. doi:10.1111/j.1467-9868.2010.00767.x.

# Multiscale Adaptive Regression Models for Neuroimaging Data

**Yimei Li**, **Hongtu Zhu**, **Dinggang Shen**, **Weili Lin**, **John H. Gilmore**, and **Joseph G Ibrahim**
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

## Summary

Neuroimaging studies aim to analyze imaging data with complex spatial patterns in a large number of locations (called voxels) on a two-dimensional (2D) surface or in a 3D volume. Conventional analyses of imaging data include two sequential steps: spatially smoothing imaging data and then independently fitting a statistical model at each voxel. However, conventional analyses suffer from the same amount of smoothing throughout the whole image, the arbitrary choice of smoothing extent, and low statistical power in detecting spatial patterns. We propose a multiscale adaptive regression model (MARM) to integrate the propagation–separation (PS) approach (Polzehl and Spokoiny, 2000, 2006) with statistical modeling at each voxel for spatial and adaptive analysis of neuroimaging data from multiple subjects. MARM has three features: being spatial, being hierarchical, and being adaptive. We use a multiscale adaptive estimation and testing procedure (MAET) to utilize imaging observations from the neighboring voxels of the current voxel to adaptively calculate parameter estimates and test statistics. Theoretically, we establish consistency and asymptotic normality of the adaptive parameter estimates and the asymptotic distribution of the adaptive test statistics. Our simulation studies and real data analysis confirm that MARM significantly outperforms conventional analyses of imaging data.

### Keywords

Kernel; Multiscale adaptive regression; Neuroimaging data; Propagation separation; Smoothing; Sphere; Test statistics

## 1. Introduction

Many large neuroimaging studies, which collect data that include anatomical and functional images from multiple subjects, have been or are being widely conducted to better understand the neural development of neuropsychiatric and neurodegenerative disorders and normal brains. By using anatomical images, various morphometrical measures of the morphology of the cortical and subcortical structures (e.g., hippocampus) are extracted to investigate neuroanatomical differences in brain structure across different populations (Thompson and Toga, 2002; Chung *et al.*, 2005). By using diffusion tensor images, various diffusion properties (e.g., fractional anisotropy) and fiber tracts are extracted for quantitative assessment of anatomical connectivity in a single subject and across different populations (Basser *et al.*, 1994; Zhu *et al.*, 2007b). Functional imaging, including functional magnetic resonance imaging (fMRI), has been widely used to understand functional integration of different brain regions in a single subject and across different populations (Friston, 2007; Huettel *et al.*, 2004).

Address for correspondence and reprints: Hongtu Zhu, Ph.D., Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420. hzhu@bios.unc.edu.

Following spatial normalization, each subject's data consists of data points at a large number of locations (called voxels), that number in the thousands to millions, on a common two-dimensional (2D) surface or in a common 3 dimensional (3D) volume. Conventional analyses of such high-dimensional imaging data are often executed by voxel-wise methods, which are carried out in two sequential steps: spatially smoothing the imaging data and then independently fitting a statistical model, such as a general linear model (LM), at each voxel. Most smoothing methods are independent of the imaging data and apply the same amount of smoothness throughout the whole image. See, for example Yue, Loh and Lindquist (2010) for overviews of smoothing methods used in the neuroimaging literature. As shown in Polzehl and Spokoiny (2000, 2006), Qiu (2005, 2007), and Tabelow *et al*. (2006, 2008a, b, c), these smoothing methods can be very problematic near the edges of the significant regions. Polzehl and Spokoiny (2000, 2006) proposed a powerful propagation–separation (PS) approach to adaptively and spatially smooth images from a single subject. Tabelow *et al*. (2006, 2008a, b, c) used the original PS idea to develop a multiscale adaptive linear model to adaptively and spatially denoise fMRI and diffusion tensor images from a single subject.

The existing voxel-wise methods for analyzing high-dimensional data involve fitting a statistical model, such as LM, to neuroimaging data from all subjects at each voxel, and then generating a statistical parametric map of test statistics and *p*-values (Lazar, 2008; Worsley *et al.*, 2004). Such methods have some obvious limitations for the analysis of neuroimaging data, which underscore the great need for further methodological development. As shown in Hecke *et al*. (2009) and Jones *et al*. (2005), voxel-wise methods can suffer from the arbitrary choice of smoothing extent in the initial smoothing step and thus dramatically increase the number of false positives and false negatives. Furthermore, as pointed out by Worsley (2003) and Tabelow *et al.* (2006), voxel-wise methods treat all voxels as independent units and do not employ the fact that the significant regions of interest have a spatial extent. Neuroimaging data, however, are spatially dependent in nature, where we often observe spatially contiguous effect regions with rather sharp edges, as is often the case in many neuroimaging analyses.

Spatially modeling neuroimaging data in the 3D volume (or 2D surface) represents both computational and theoretical challenges. It is common to use conditional autoregressive (CAR), Markov random field (MRF), and other spatial correlation priors to characterize spatial dependence among spatially connected voxels (Besag, 1986; Banerjee, Carlin, and Gelfand, 2004). However, calculating the normalizing factor of MRF and estimating spatial correlation for a large number of voxels in the 3D volume (or 2D surface) are computationally prohibitive (Zhu, Gu, and Peterson, 2007; Bowman, 2007). Moreover, it can be restrictive to assume a specific type of correlation structure, such as CAR and MRF, for the whole 3D volume (or 2D surface).

The goal of this article is to develop a multiscale adaptive regression model (MARM) for the spatial and adaptive analysis of neuroimaging data. MARM integrates the PS approach and voxel-wise methods and thus it is a generalization of the PS approach (Polzehl and Spokoiny, 2000, 2006) to neuroimaging data from multiple subjects. MARM has three features: being spatial, being hierarchical and being adaptive. MARM can efficiently combine all observations with adaptive weights in the voxels within the sphere of the current voxel to increase the precision of parameter estimates and the power of test statistics in detecting subtle changes of brain structure and function. Due to its hierarchical and adaptive nature, MARM can efficiently learn the shape of activation areas, use the adaptive weights to capture shape information, and then preserve the edges of activation areas.

MARM provides a general probability framework for adaptively carrying out statistical inference on neuroimaging data obtained from multiple subjects. We establish consistency and asymptotic normality of the adaptive estimator and the asymptotic distribution of the adaptive test statistic for MARM as the number of subjects (or images) increases to infinity. The covariance estimate of the adaptive estimator in MARM has a simple form. Our new theoretical results show that in MARM, the adaptive weighting idea of the novel PS approach is valid without imposing the propagation condition. Our results show that it is critical to choose appropriate parameters in constructing adaptive weights in order to have simple asymptotic results to carry out statistical inference including hypothesis testing.

To motivate the proposed methodology, we consider fractional anisotropy (FA) imaging data acquired at 2 weeks, year 1 and year 2 from 38 subjects in a neonatal project on early brain development, which is discussed in more detail in Section 4. The primary interest here was to identify the spatial patterns of white matter maturation. We smoothed FA imaging data with two levels of smoothness. Then, at each voxel, we fit a multivariate linear model with age and age$^2$ as covariates and calculated the Wald statistics and their associated $p$ values for testing age dependent effect. Inspecting Figure 3 reveals that the size of significant regions and degree of significance associated with the age dependent effect strongly depend on the size of smoothness, which agrees with the findings in Jones *et al*. (2005). We also analyzed the same FA dataset using MARM and tested the age dependent effect across all voxels. MARM can preserve the edges of significant regions compared with the results from the smoothed images (Figs. 3(b)–(c)). In contrast, the significant regions based on the smoothed images even spread over cerebrospinal fluid (CSF) areas (Fig. 3(c)), in which FA values should be close to zero and have no age dependent effect. In Section 4, we will revisit this data set.

Section 2 of this paper presents MARM and establishes the associated theoretical properties. We establish consistency and asymptotic normality of the adaptive estimator and the asymptotic distribution of the adaptive test statistic for MARM. In Section 3, we conduct simulation studies with the known ground truth to examine the finite sample performance of the adaptive estimates and test statistics in MARM. Section 4 illustrates an application of the proposed methods in a real neuroimaging dataset. We present concluding remarks in Section 5.

## 2. Multiscale Adaptive Regression Model

### 2.1. Model Formulation

We consider imaging measurements in the 3D volume (or on the 2D surface) and clinical variables from $n$ subjects. Without loss of generality, we focus on the 3D volume. Let $\mathcal{D}$ and $d$, respectively, represent a 3D volume and a voxel in $\mathcal{D}$, $m$ be an integer, and $N(\mathcal{D})$ equal the number of voxels in $\mathcal{D}$. For the $i$th subject, we observe an $m \times 1$ vector of imaging measures $Y_i(d)$ at voxel $d$, which leads to an $mN(\mathcal{D}) \times 1$ vector of measurements across $\mathcal{D}$, denoted by $\mathbf{Y}_{i,\mathcal{D}} = \{Y_i(d) : d \in \mathcal{D}\}$, and a $p_1 \times 1$ vector of clinical variables $\mathbf{x}_i$. In neuroimaging studies, imaging measurements can include the shape representation of the surfaces of cortical or subcortical structures, fMRI signals, diffusion tensors, and so on (Ashburner and Friston, 2000; Thompson and Toga, 2002). Clinical variables often include pedigree information, time, demographic characteristics (e.g., age, gender, and height), and diagnostic status among others.

Statistically, our primary interest is to build the conditional distribution of $\mathbf{Y}_{\mathcal{D}} = \{\mathbf{Y}_{i,\mathcal{D}} : i = 1, \cdots, n\}$ given $\mathbf{X} = \{\mathbf{x}_i : i = 1, \cdots, n\}$, that is, $p(\mathbf{Y}_{\mathcal{D}}|\mathbf{X})$. For a cross-sectional design, it is natural to assume that data from different subjects are independent, that is $p(\mathbf{Y}_{\mathcal{D}}|\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$.

Thus, we only need to specify $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$ for each $i$. However, the number of voxels in each brain region can be more than 500,000 voxels, and at each voxel, the dimension of $Y_i(d)$ can be univariate or multivariate, thus totaling a billion or more data points in an entire study. In addition, imaging data $\mathbf{Y}_{i,\mathcal{D}}$ are spatially dependent in nature, and thus given the large number of voxels on each brain structure, it is statistically challenging to model the spatial relationships among all pairs of points simultaneously.

The voxel-wise approach essentially assumes that

$$p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)=\prod_{d\in\mathcal{D}} p(Y_i(d)|\mathbf{x}_i,\theta(d)), \tag{1}$$

where $p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d))$ is the marginal density of $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$ at voxel $d$ and $\boldsymbol{\theta}(d) = (\theta_1(d), \cdots, \theta_p(d))^T$ is a $p \times 1$ vector in an open subset $\Theta$ of $R^p$, in which $p$ is an integer. Moreover, the voxel-wise approach makes a strong perfect registration assumption. That is, after an image warping procedure, the location of a voxel in the images of one person is assumed to be in precisely the same location as the voxel identified in another person. Note that due to possible model misspecification, $p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d))$ is only a 'pseudo' density function for $Y_i(d)$. Model (1) is general enough to comprise most statistical models including LM's in the neuroimaging literature. However, since the voxel-wise approach does not account for the spatial nature of neuroimaging data, which often shows effects in spatially contiguous regions with rather sharp edges, it may lead to a loss of power in detecting statistical significance in the analysis of neuroimaging data.

To utilize the spatial nature of neuroimaging data, the multiscale adaptive regression model is developed as follows. In many neuroimaging studies, our primary interest is to make statistical inference about $\boldsymbol{\theta}(d)$ at each voxel $d \in \mathcal{D}$. Instead of solely using the data in voxel $d$, it would be more efficient to utilize all the data in the neighboring voxels of $d$ to estimate $\boldsymbol{\theta}(d)$. Similar to standard kernel smoothing methods (Qiu, 2006), we consider a spherical neighborhood of $d$ with a radius (or bandwidth) $r_0$, denoted by $B(d, r_0)$. By assuming spatial independence among $\{Y_i(d') : d' \in B(d, r_0)\}$, we construct a weighted likelihood to estimate $\boldsymbol{\theta}(d)$, denoted by $p_W(Y_i(d') : d' \in B(d, r_0)|\mathbf{x}_i, \boldsymbol{\theta}(d))$, as follows:

$$p_W(Y_i(d'):d' \in B(d,r_0)|\mathbf{x}_i,\theta(d))=\prod_{d'\in B(d,r_0)} p(Y_i(d')|\mathbf{x}_i,\theta(d))^{\omega(d,d';r_0)}, \tag{2}$$

where $\omega(d, d'; h)$ characterizes the similarity between the data in voxels $d'$ and $d$ with $\omega(d, d; h) = 1$. If $\omega(d, d'; h) \approx 0$, then $p(Y_i(d')|\mathbf{x}_i, \boldsymbol{\theta}(d))^{\omega(d,d',;r_0)}$ is close to 1 and thus the observations in voxel $d'$ do not provide information on $\boldsymbol{\theta}(d)$. Therefore, $\omega(d, d'; r_0)$ can prevent incorporation of voxels whose data do not contain information on $\boldsymbol{\theta}(d)$ and preserve the edges of significant regions. In neuroimaging data, voxels, which are not on the boundary of regions of significance (Fig. 1(c)), often have a neighborhood in which $\boldsymbol{\theta}(d)$ is nearly constant. In this case, $\omega(d, d'; h)$ for voxel $d'$ in the neighborhood of voxel $d$ is greater than zero and thus $p_W(\boldsymbol{\theta}(d)|Y_i(d') : d' \in B(d, r_0))$ allows borrowing 'good' information from these neighboring voxels. Furthermore, we assume that $\omega(d, d'; h)$ is independent of $i$ just for notational simplicity.

Let $\boldsymbol{\omega} = \{\omega(d, d'; r_0) : d \in \mathcal{D}, d' \in B(d, r_0)\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}(d) : d \in \mathcal{D}\}$. Finally, by assuming the spatial independence among imaging data, we take the product of $p_W(Y_i(d') : d' \in B(d, r_0)|\mathbf{x}_i, \boldsymbol{\theta}(d))$ for all $d \in \mathcal{D}$ and then obtain a weighted likelihood function of MARM for $\mathbf{Y}_{i,\mathcal{D}}$ given by

$$p_W(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i,\theta,\omega)=\prod_{d\in\mathcal{D}}\{\prod_{d'\in B(d,r_0)}p(Y_i(d')|\mathbf{x}_i,\theta(d))^{\omega(d,d';r_0)}\}. \tag{3}$$

When $r_0 = 0$, $B(d, r_0)$ and model (3), respectively, reduce to $d$ and model (1) for the voxel-wise method.

## 2.2. Examples

MARM can be applied to the analysis of neuroimaging data from multiple subjects and those from a single subject. For the case of a single subject, MARM reduces to the PS approach. For the purposes of illustration, we consider the following three examples.

**Example 1—**We consider a multivariate nonlinear model at each voxel given by

$$Y_i(d)=\mu(\mathbf{x}_i,\beta(d))+\boldsymbol{\varepsilon}_i(d) \tag{4}$$

for $i = 1, \cdots, n$ and $d \in \mathcal{D}$, where $\mu(\cdot, \cdot)$ is a known $m \times 1$ vector of nonlinear functions, $\boldsymbol{\beta}(d)$ is a $p_2 \times 1$ vector representing unknown regression coefficients, and $\boldsymbol{\varepsilon}_i(d)$ is an $m \times 1$ random vector with mean zero and covariance matrix $\Sigma(d)$. In this case, $\boldsymbol{\theta}(d)$ contains all parameters in $\boldsymbol{\beta}(d)$ and $\Sigma(d)$. If we use the density of the Gaussian distribution to approximate $p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d))$ and assume the spatial independence among imaging data, then $\log p_W(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\omega})$ based on model (4) is given by

$$-\sum_{d\in\mathcal{D}}\sum_{d'\in B(d,r_0)}0.5\omega(d,d';r_0)[\log|\sum(d)|+\{Y_i(d')-\mu(\mathbf{x}_i,\beta(d))\}^T\sum(d')^{-1}\{Y_i(d')-\mu(\mathbf{x}_i,\beta(d))\}]. \tag{5}$$

If $\mu(\mathbf{x}_i, \boldsymbol{\beta}(d)) = X_i\boldsymbol{\beta}(d)$, where $X_i$ is an $m \times p_2$ covariate matrix of $\mathbf{x}_i$, then model (4) reduces to the multi-scale adaptive multivariate linear model for the analysis of neuroimaging data (Tabelow *et al*. 2006Tabelow *et al*. 2008a,b,c).

**Example 2—**We consider a generalized linear model (GLM) for the conditional distribution of $Y_i(d)$ given $\mathbf{x}_i$ (McCullagh and Nelder 1989). Specifically, for $i = 1, \ldots, n$, $Y_i(d)$ given $\mathbf{x}_i$ has a density in the exponential family

$$\exp(\tau(d)\{Y_i(d)\eta_i(\beta(d))-b(\eta_i(\beta(d)))\}+c(Y_i(d),\tau(d))], \tag{6}$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions. Moreover, $\eta_i(\beta(d))=\eta(g(\mathbf{x}_i^T\beta(d)))$ for $i = 1, \ldots, n$, where $g(\cdot)$ is a known and monotonic link function and $\boldsymbol{\beta}(d)$ is a $(p - 1) \times 1$ vector of regression coefficients. In this case, $\boldsymbol{\theta}(d) = (\boldsymbol{\beta}(d), \tau(d))$ and the weighted quasi-likelihood function of MARM under spatial independence is given by

$$\sum_{i=1}^{n}\sum_{d\in\mathcal{D}}\sum_{d'\in B(d,r_0)}\omega(d,d';r_0)[\tau(d)\{Y_i(d')\eta_i(\beta(d))-b(\eta_i(\beta(d)))\}+c(Y_i(d'),\tau(d))]. \tag{7}$$

**Example 3—**In a fMRI session, $n$ fMRI volumes are acquired at acquisition times $t_1, \cdots, t_n$ while a subject performs a cognitive or behavioral task. At each voxel, we consider a regression model $Y_i(d) = \mu(\mathbf{x}_i, \beta(d)) + \varepsilon_i(d)$, where $\varepsilon_i(d)$ denotes measurement errors with mean zero and variance $1/\tau(d)$ and $\mathbf{x}_i$ may include responses to differing stimulus types, the rest status, and various reference functions (Lazar, 2008; Tabelow *et al.*, 2006, 2008a, c). The measurement errors $\varepsilon_i(d)$ may include noise from stochastic variation, numerous physiological processes, eddy currents, artifacts from the differing magnetic field susceptibilities of neighboring tissues, non-rigid motion, preprocessing methods (registration, normalization) among many others (Huettel *et al.*, 2004; Lazar, 2008). By performing a prewhitening procedure, we may assume that $\{\varepsilon_i(d) : i = 1, \cdots, n\}$ have zero mean and are approximately uncorrelated. If we use the density of the Gaussian distribution to approximate $p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d))$, where $\boldsymbol{\theta}(d) = (\boldsymbol{\beta}(d), \tau(d))$, then the weighted quasi-likelihood function of MARM for fMRI is given by

$$\sum_{i=1}^{n}\sum_{d\in\mathcal{D}}\sum_{d'\in B(d,r_0)} 0.5\omega(d,d';r_0)[\log\tau(d) - \tau(d)\{Y_i(d') - \mu(\mathbf{x}_i,\beta(d))\}^2].$$

## 2.3. Multiscale Adaptive Estimation and Testing Procedure

We use a multiscale adaptive estimation and testing (MAET) procedure to determine $\boldsymbol{\omega}$, estimate $\boldsymbol{\theta}(d)$, and calculate its associated test statistic across all voxels. MAET uses the same multiscale adaptive strategy from the PS approach (Polzehl and Spokoiny, 2000, 2006), and thus it can be regarded as a generalization of the PS approach to neuroimaging data with multiple subjects. MAET starts with building a sequence of nested spheres with increasing radii $h_0 = 0 < h_1 < \cdots < h_S = r_0$ ranging from the smallest scale $h_0 = 0$ to the largest scale $h_S = r_0$ at each $d \in \mathcal{D}$ (panel (b) in Fig. 1). By setting $\omega(d, d'; h_0) = 1$, we can estimate $\boldsymbol{\theta}(d)$ at scale $h_0$, denoted by $\hat{\boldsymbol{\theta}}(d; h_0)$, and construct a test statistic $W_\mu(d, h_0)$. Then, based on the information contained in $\{\hat{\boldsymbol{\theta}}(d; h_0) : d \in \mathcal{D}\}$, we use methods as detailed below to calculate weights $\omega(d, d'; h_1)$ at scale $h_1$ for all $d \in \mathcal{D}$. In this way, we can sequentially determine $\omega(d, d'; h_s)$ and adaptively update $\hat{\boldsymbol{\theta}}(d; h_s)$ and $W_\mu(d, h_s)$, which are defined in (9) and (12), respectively, as the radius ranges from $h_0 = 0$ to $h_S = r_0$.

Specifically, for a given radius, we consider maximum weighted likelihood estimates of $\boldsymbol{\theta}(d)$ across all voxels $d \in \mathcal{D}$ given the current fixed weights $\{\omega(d, d'; h) : d, d' \in \mathcal{D}\}$. Let $\tilde{\omega}(d, d'; h) = \omega(d, d'; h)/\Sigma_{d' \in B(d,h)}\, \omega(d, d'; h)$. For the sphere with radius $h$ of the voxel $d$, based on model (3), we consider a normalized weighted quasi-likelihood function $\ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}})$, which is given by

$$\ell_n(\theta(d);h,\tilde{\omega})=\sum_{i=1}^{n}\sum_{d'\in B(d,h)}\tilde{\omega}(d,d';h)\log p(Y_i(d')|\mathbf{x}_i,\theta(d)). \tag{8}$$

The $\ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}})$ utilizes all the data in $\{Y_i(d') : d' \in B(d, h)\}$ and normalized weights $\{\omega(d, d'; h) : d' \in B(d, h)\}$. The maximum weighted quasi-likelihood (MWQL) estimate of $\boldsymbol{\theta}(d)$, denoted by $\hat{\boldsymbol{\theta}}(d, h)$, is defined by

$$\widehat{\theta}(d,h)=\operatorname{argmax}_{\theta(d)} n^{-1}\ell_n(\theta(d);h,\tilde{\omega}). \tag{9}$$

Numerically, we use various optimization algorithms, such as a Newton-Raphson type algorithm, to estimate $\hat{\boldsymbol{\theta}}(d, h)$. After convergence, $\mathrm{Cov}(\hat{\boldsymbol{\theta}}(d, h))$ can be approximated by

$$\mathrm{Cov}(\widehat{\theta}(d,h)) \approx \sum\nolimits_n (\widehat{\theta}(d,h)) = \{\sum\nolimits_{n,1} (\widehat{\theta}(d,h))\}^{-1} \sum\nolimits_{n,2} (\widehat{\theta}(d,h)) \{\sum\nolimits_{n,1} (\widehat{\theta}(d,h))\}^{-1}, \tag{10}$$

where $\sum_{n,1}(\theta(d)) = -\partial^2_{\theta(d)} \ell_n(\theta(d); h, \tilde{\omega})$ and $\sum_{n,2}(\theta(d)) = \sum_{i=1}^{n} \{\sum_{d' \in B(d,h)} \tilde{\omega}(d, d'; h) \partial_{\theta(d)} \log p(Y_i(d') | \mathbf{x}_i, \theta(d))\}^{\otimes 2}$, in which $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for any vector $\mathbf{a}$.

Our choice of which hypotheses to test is motivated by either a comparison of brain structure (or function) across diagnostic groups or the detection of a change in brain structure (or function) across time (Chung *et al.*, 2005; Lazar, 2008; Thompson and Toga, 2002). These questions of interest usually can be formulated as testing hypotheses about $\boldsymbol{\theta}(d)$ as follows:

$$H_{0,\mu}: R(\theta(d)) = \mathbf{b}_0 \quad \text{vs.} \quad H_{1,\mu}: R(\theta(d)) \neq \mathbf{b}_0, \tag{11}$$

where $R(\boldsymbol{\theta}(d))$ is an $r \times 1$ vector function of $\boldsymbol{\theta}(d)$ with $p \geq r$ and $\mathbf{b}_0$ is an $r \times 1$ specified vector, such as an $r \times 1$ vector of zeros. We test the null hypothesis $H_{0,\mu}$ using the Wald test statistic $W_\mu(d, h)$, which is given by

$$W_\mu(d,h) = \{R(\widehat{\theta}(d;h)) - \mathbf{b}_0\}^T \{\partial_{\theta(d)} R(\widehat{\theta}(d;h)) \widehat{\sum}_n (\widehat{\theta}(d;h)) \partial_{\theta(d)} R(\widehat{\theta}(d;h))^T\}^{-1} \{R(\widehat{\theta}(d;h)) - \mathbf{b}_0\}. \tag{12}$$

A path diagram of MAET is given below:

$$\begin{array}{ccccccc} \omega(d,d';h_0) & & \omega(d,d';h_1) & & \cdots & & \omega(d,d';h_s = r_0) \\ \Downarrow & \nearrow & \Downarrow & \nearrow & \cdots & \nearrow & \Downarrow \\ \widehat{\theta}(d;h_0) & & \widehat{\theta}(d;h_1) & & \cdots & & (\widehat{\theta}(d;h_s), W_\mu(d;h_s)). \end{array} \tag{13}$$

At each iteration, the computations involved for MARM are of the same order as that for the voxel-wise approach. Thus, this multiscale adaptive method provides an efficient method for flexibly exploring the neighboring areas of each voxel. Since MARM sequentially includes more data at each iteration, it will adaptively increase the statistical efficiency in estimating $\boldsymbol{\theta}(d)$ in a homogenous region and decrease the variation of the weights $\omega(d, d'; h)$.

The MAET procedure consists of five key steps: (i) initialization, (ii) weights adaptation, (iii) estimation, (iv) stop checking, and (v) inference. In the initialization step (i), we fix a geometric series $\{h_s = c_h^s : s = 1, \ldots, S\}$ of radii with $h_0 = 0$, where $c_h > 1$, say $c_h = 1.10$. The parameter $c_h^s$ plays the same role as the bandwidth of local kernel methods. A small value of $c_h$ only allows incorporating the closest neighboring voxels and thus it can prevent oversmoothing $\boldsymbol{\theta}(d)$ at the beginning of MAET, whereas a small $c_h$ leads to increased computational effort. At each voxel $d$, let $\omega(d, d'; h_0) = \mathbf{1}(d = d')$, in which $\mathbf{1}(\cdot)$ is an indicator function. Then, we calculate the MWQL estimate $\hat{\boldsymbol{\theta}}(d, h_0)$, which is defined in (9) at each voxel $d \in \mathcal{D}$. The $\hat{\boldsymbol{\theta}}(d, h_0)$ are the same as those from the voxel-wise approach. We then set $s = 1$ and $h_1 = c_h$.

In the weight adaptation step (ii), we compute the similarity between voxels $d$ and $d'$, denoted by $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$, and the adaptive weights $\omega(d, d'; h_s)$, which are, respectively, defined as

$$D_\theta(d,d';h_{s-1})=\{\widehat{\theta}(d,h_{s-1})-\widehat{\theta}(d',h_{s-1})\}^T\widehat{\sum}(\widehat{\theta}(d;h_{s-1}))^{-1}\{\widehat{\theta}(d,h_{s-1})-\widehat{\theta}(d',h_{s-1})\}, \tag{14}$$

$$\omega(d,d';h_s)=K_{loc}(\|d-d'\|_2/h_s)K_{st}(D_\theta(d,d';h_{s-1})/C_n), \tag{15}$$

where $K_{loc}(u)$ and $K_{st}(u)$ are two nonnegative kernel functions with compact support such that all of them decrease to zero as $u$ increases, $C_n$ is a number, which may be associated with $n$, and $\|\cdot\|_2$ denotes the Euclidean norm of a vector (or a matrix). The weights $K_{loc}(\|d - d'\|_2/h_s)$ give less weight to the voxel $d' \in B(d, h_s)$, whose location is far from the voxel $d$. The weights $K_{st}(u)$ downweight the voxels $d'$ with large $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$, which indicates a large difference between $\hat{\boldsymbol{\theta}}(d', h_{s-1})$ and $\hat{\boldsymbol{\theta}}(d, h_{s-1})$.

In the estimation step (iii), for the radius $h_s$, we substitute $\omega(d, d'; h_s)$ into (9) to calculate $\hat{\boldsymbol{\theta}}(d, h_s)$ and then compute $W_\mu(d, h_s)$ according to (12) at each voxel $d \in \mathcal{D}$.

In the stop checking step (iv), after the $S_0$-th iteration, we calculate a stopping criterion based on a normalized distance between $\hat{\boldsymbol{\theta}}(d; h_{S_0})$ and $\hat{\boldsymbol{\theta}}(d; h_s)$ for $s > S_0$, which is given by

$$D(\widehat{\theta}(d;h_{s_0}),\widehat{\theta}(d;h_s))=\{\widehat{\theta}(d,h_{s_0})-\widehat{\theta}(d,h_s)\}^T\widehat{\sum}(\widehat{\theta}(d;h_{s_0}))^{-1}\{\widehat{\theta}(d,h_{s_0})-\widehat{\theta}(d,h_s)\}. \tag{16}$$

Then, we check whether $\hat{\boldsymbol{\theta}}(d; h_s)$ is in an $\alpha$ confidence ellipsoid of $\hat{\boldsymbol{\theta}}(d; h_{S_0})$ given by $\{\boldsymbol{\theta}: D(\hat{\boldsymbol{\theta}}(d; h_{S_0}), \boldsymbol{\theta}(d)) \le \tilde{C} = \chi^2(p)^\alpha\}$, where $\chi^2(p)^b$ is the upper $1-b$ percentile of the $\chi^2(p)$ distribution. To prevent a large $D(\hat{\boldsymbol{\theta}}(d; h_{S_0}), \hat{\boldsymbol{\theta}}(d; h_s))$, we set $\alpha = 80\%$ in the paper. If $D(\hat{\boldsymbol{\theta}}(d; h_{S_0}), \hat{\boldsymbol{\theta}}(d; h_s))$ is greater than $\tilde{C}$, then we set $\hat{\boldsymbol{\theta}}(d, h_S) = \hat{\boldsymbol{\theta}}(d, h_{s-1})$, $W_\mu(d, h_S) = W_\mu(d, h_{s-1})$, and $s = S$. If $s = S$, we go to the inference step (v). If $s \le S_0$ or $D(\hat{\boldsymbol{\theta}}(d; h_{S_0}), \hat{\boldsymbol{\theta}}(d; h_s)) \le C$ for $S - 1 \ge s > S_0$, then we set $h_{s+1} = c_h h_s$, increase $s$ by 1 and continue with the weight adaptation step (ii).

In the inference step (v), when $s = S$, we report the final $\hat{\boldsymbol{\theta}}(d, h_S)$, compute the $p$-values for $W_\mu(d, h_S)$, correct for multiple comparisons by using either the Bonferroni correction, the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) or random field theory (Worsley *et al.*, 2004; Nichols and Hayasaka, 2003), and then stop the algorithm.

**Example 4—**As an illustration, we consider the multiscale adaptive multivariate linear model described in Example 1 and present the key components of the four steps of MAET as follows. In the initialization step (i), at each voxel $d$, by setting $\hat{\Sigma}(d, h_0)^{(0)} = \mathbf{I}_m$, an $m \times m$ identity matrix, we iteratively update

$$\begin{aligned}\widehat{\beta}(d,h_0)^{(t+1)}&=[\sum_{i=1}^n X_i^T\{\widehat{\sum}(d,h_0)^{(t)}\}^{-1}X_i]^{-1}\sum_{i=1}^n X_i^T\{\widehat{\sum}(d,h_0)^{(t)}\}^{-1}Y_i(d),\\ \widehat{\sum}(d,h_0)^{(t+1)}&=(n-p_1)^{-1}\sum_{i=1}^n\{Y_i(d)-X_i\widehat{\beta}(d,h_0)^{(t+1)}\}^{\otimes 2}\end{aligned} \tag{17}$$

until convergence. Since in most neuroimaging applications, $\boldsymbol{\beta}$ is the primary parameter of interest, we fix $\Sigma(d)$ at $\hat{\Sigma}(d, h_0)$ at each $d$. Then, we compute

$$\mathrm{Cov}(\widehat{\beta}(d,h_0)) \approx \sum\nolimits_n (\widehat{\beta}(d,h_0))=\{\sum\nolimits_{i=1}^{n} X_i^T \widehat{\sum}(d,h_0)^{-1} X_i\}^{-1}.$$

In the weight adaption step (ii), compute $D_\beta(d, d'; h_{s-1}) = \{\hat{\boldsymbol{\beta}}(d, h_{s-1}) - \hat{\boldsymbol{\beta}}(d', h_{s-1})\}^T \Sigma_n(\hat{\boldsymbol{\beta}}(d; h_{s-1}))^{-1}\{\hat{\boldsymbol{\beta}}(d, h_{s-1}) - \hat{\boldsymbol{\beta}}(d', h_{s-1})\}$ and $\omega(d, d'; h_s) = K_{loc}(\|d - d'\|_2/h_s)K_{st}(D\boldsymbol{\beta}(d, d'; h_{s-1})/C_n)$.

In the estimation step (iii), for the radius, let

$$A(d,h_s,\omega;X)=\sum\nolimits_{i=1}^{n} X_i^T \sum\nolimits_{d' \in B(d,h_s)} \omega(d,d';h_s)\widehat{\sum}(d',h_0)^{-1} X_i, \text{ compute}$$

$$\widehat{\beta}(d,h_s)=A(d,h_s,\omega;X)^{-1}[\sum\nolimits_{i=1}^{n} X_i^T \sum\nolimits_{d' \in B(d,h_s)} \omega(d,d';h_s)\widehat{\sum}(d',h_0)^{-1} Y_i(d')] \text{ and}$$

$$\sum\nolimits_n (\widehat{\beta}(d,h_s))=A(d,h_s,\omega;X)^{-1}\{\sum_{i=1}^{n} X_i^T \widehat{\varepsilon}_i(d;\omega,h_s)^{\otimes 2} X_i\}A(d,h_s,\omega;X)^{-1},$$

where $\hat{\boldsymbol{\varepsilon}}_i(d; \boldsymbol{\omega}, h_s) = \Sigma_{d' \in B(d,hs)}\, \omega(d, d'; h_s)\hat{\Sigma}(d', h_0)^{-1}\{Y_i(d') - X_i\hat{\boldsymbol{\beta}}(d', h_s)\}$.

In the stop checking step (iv), we compute $D(\hat{\boldsymbol{\beta}}(d; h_{S_0}), \hat{\boldsymbol{\beta}}(d; h_s)) = \{\hat{\boldsymbol{\beta}}(d, h_{S_0}) - \hat{\boldsymbol{\beta}}(d, h_s)\}^T \hat{\Sigma}(\hat{\boldsymbol{\beta}}(d; h_{S_0}))^{-1}\{\hat{\boldsymbol{\beta}}(d, h_{S_0}) - \hat{\boldsymbol{\beta}}(d, h_s)\}$ for $s > S_0$.

### 2.4. Parameters of MAET

The performance of MAET depends on specifying the following parameters of MAET, including $c_h$, $C_n$, $K_{loc}(u)$, $K_{st}(u)$, $S_0$, and $S$. We have tested different combinations of these parameters of MAET in both simulated and real imaging data. According to our experience, the performance of MAET is quite robust to moderate changes in these parameters.

We suggest choosing a relatively small $c_h$. The $c_h$ is essentially the bandwidth of local kernel methods. When voxel $d$ is near/on the edge of regions with distinct features, $B(d, c_h)$ for a large $c_h$ may include voxels from these distinct regions, which can cause oversmoothing of the parameter estimates image. In contrast, even when voxel $d$ is near, but not on the edge of distinct regions, $B(d, c_h)$ for small $c_h$ only includes the closest neighboring voxels $d'$, whose data are similar to those of voxel $d$, and thus it can improve the accuracy of parameter estimation in the first few iterations. Subsequently, when combined with the stop checking step, small $c_h$ can improve the robustness of MAET and the accuracy of parameter estimation across all voxels.

The $C_n$ is used to penalize the similarity between any two voxels $d$ and $d'$. If there is a moderate similarity between the voxels $d$ and $d'$, a large $C_n$ leads to small $D(d, d'; h_s)/C_n$ and thus it decreases the sensitivity of MAET in separating such voxels. Thus, a large $C_n$ can increase the estimation error near the boundary of two regions with distinct features, when the difference between the two regions is moderate. In contrast, when voxels $d$ and $d'$ are similar to each other with a small $D(d, d'; h_s)$, a small $C_n$ may lead to a relative large $D(d, d'; h_s)/C_n$ and thus it may decrease the specificity of MAET in combining such similar voxels. Thus, a small $C_n$ can decrease the accuracy of parameter estimation in the interior of a homogeneous region. Therefore, a good $C_n$ should balance between the sensitivity and specificity of MAET. So far, we have tested various values of $C_n$ by using simulation studies, among which $n^{0.4}\chi^2(p)^{0.95}$ and $\log(n)\chi^2(p)^{0.95}$ perform equally well. Without loss of generality, we set $\log(n)\chi^2(p)^{0.95}$. However, to account for the variability in estimating $\Sigma_n(\hat{\boldsymbol{\theta}}$

$(d, h_S)$), it may be more suitable to use the quantiles of the $F$ distribution instead of the $\chi^2$ distribution.

The $K_{loc}(u)$ is a regular kernel function for further smoothing curves or surfaces based on the Euclidean distance between voxels. Some common choices of $K_{loc}(u)$ include the Gaussian kernel and the Epanechnikov kernel (Tabelow *et al.*, 2006Tabelow *et al.*, 2008a, b, c; Polzehl and Spokoiny, 2000, 2006). Because MAET mainly uses the similarity information between any pairs of voxels, the specification of $K_{loc}(u)$ is not critical for MAET. We use $K_{loc}(u) = (1 - u)_+$.

We set $K_{st}(u) = \exp(-u)$ in our simulated and real imaging data. Theoretically, as shown later, $\exp(-u)$ gives an exponential decay rate of $n$. Although different choices of $K_{st}(\cdot)$ have been suggested in the original PS approach (Polzehl and Spokoiny, 2000, 2006; Tabelow *et al.*, 2006, 2008a, b, c), 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 $\min(1, 2(1 - u))_+$, which has better performance in spatially and adaptively smoothing fMRI and DTI from a single subject (Polzehl and Tabelow, 2007).

We suggest not to set $S_0$ as 0 or a large integer. If $S_0 = 0$, then only the data in voxel $d$ are included and the accuracy of $\hat{\boldsymbol{\theta}}(d, h_0)$ may be low. For large $S_0$, since the number of voxels in $B(d, h_{S_0})$ is large, it easily leads to both heavy computation and oversmoothing when voxel $d$ is either on the boundary of significant regions or in some regions in which the parameters change slowly with voxel location. After the $S_0$-th iteration, the stop checking step starts to compute the stopping criterion and check whether further iteration is needed in this voxel. Since $c_h^s$ plays the same role as the bandwidth in the local kernel method, the stop checking step is essentially a bandwidth selection procedure. This step is to compare consecutive parameter estimates in order to prevent bad data from neighboring voxels and oversmoothing the parameter estimates image. We have found that $S_0 = 3$ coupled with a small $c_h = 1.1$ performs very well in numerous simulations.

As the maximal iteration $S$ increases, the number of neighboring voxels in $B(d, h_s = c_h^S)$ increases exponentially. Moreover, a large $S$ also increases the probability of oversmoothing $\boldsymbol{\theta}(d)$ when the current voxel $d$ is near the edge of distinct regions and the parameters change slowly with other locations. In practice, we suggest the maximal step $S$ to be between 10 and 20.

Setting the starting value of $\hat{\boldsymbol{\theta}}(d, h_s)^{(0)}$ as $\hat{\boldsymbol{\theta}}(d, h_{s-1})$ for each $s > 0$ is an efficient way of selecting the initial value in the Newton-Raphson algorithm. Since the MAET procedure always downweights voxel $d' \in B(d, h)$ in $\ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}})$ when the value of $D_{\boldsymbol{\theta}}(d, d'; hs_{-1})$ is large, $\hat{\boldsymbol{\theta}}(d, h_{s-1})$ and $\hat{\boldsymbol{\theta}}(d, h_s)$ should be close to each other. By starting from $\hat{\boldsymbol{\theta}}(d, h_s)^{(0)} = \hat{\boldsymbol{\theta}}(d, h_{s-1})$, the Newton-Raphson algorithm converges very fast. The additional computational time for MARM is moderate compared to the voxel-wise approach, since MARM only involves some additional operation for locally averaging over all voxels in $B(d, h_s)$ at each voxel $d$.

### 2.5. Theoretical Properties

We establish the asymptotic properties of adaptive estimators and test statistics for MAET with stochastic adaptive weights. A critical question is that what kinds of stochastic weights can automatically incorporate 'good' information and prevent 'bad' information from neighboring voxels. By appropriately utilizing information from neighboring voxels, the MAET procedure can dramatically increase the accuracy and efficiency in estimating the true value $\boldsymbol{\theta}_*(d)$ in each voxel. Another important question is whether the stochastic weights

chosen can ensure consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}(d, h)$ at each fixed scale $h$. To have a better understanding of the MAET procedure, we focus on the asymptotic behavior of the adaptive weight when $s = 1$ and then discuss the scenario when $s > 1$.

Throughout the paper, we only consider the asymptotic properties of $\hat{\boldsymbol{\theta}}(d, h_s)$ and $W_\mu(d, h_s)$ for a finite number of iterations and bounded $r_0$ for MAET, since a brain volume is always bounded. We assume that the number of voxels in the brain volume does not increase with the sample size, since the resolution of a given imaging dataset is always fixed. We obtain the following theorems, whose detailed assumptions and proofs can be found in a supplementary report.

**Theorem 1—**If assumptions (C1)–(C7) in the supplementary report are true, then we have

- **a.** $\hat{\boldsymbol{\theta}}(d, h_0)$ converges to $\boldsymbol{\theta}_*(d)$ in probability;
- **b.** $\{\Sigma_{n,2}(\hat{\boldsymbol{\theta}}(d, h_0))\}^{-1/2}\Sigma_{n,1}(\hat{\boldsymbol{\theta}}(d, h_0))\{\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)\} \to^{L} N(0, \mathbf{I}_p)$, where $\to^{L}$ denotes convergence in distribution;
- **c.** $D_{\boldsymbol{\theta}}(d, d'; h_0)$ and $K_{st}(D_\theta(d, d';h_0)C_n^{-1})$ can be, respectively, approximated by

$$\begin{aligned} D_\theta(d, d';h_0)&=\mathbf{1}(\Delta_*(d, d')=\mathbf{0}) \times O_p(\log(N(\mathcal{D})))+\mathbf{1}(\Delta_*(d, d') \neq \mathbf{0}) \times n\left\|\{\Sigma_*(d)\}^{-1/2}\{\Delta_*(d, d')+O_p(\sqrt{\log(N(\mathcal{D}))/n})\}\right\|_2^2, \\ K_{st}(D_\theta(d, d';h_0)C_n^{-1})&=\mathbf{1}(\Delta_*(d, d') \neq \mathbf{0})K_{st}(C_n^{-1}nO_p(1))+\mathbf{1}(\Delta_*(d, d')=\mathbf{0})K_{st}(\log(N(\mathcal{D}))C_n^{-1}O_p(1)), \end{aligned} \tag{18}$$

  where $\Delta_*(d, d') = \boldsymbol{\theta}_*(d) - \boldsymbol{\theta}_*(d')$ and $\Sigma_*(d) = \Sigma_{1*}(d)^{-1}\Sigma_{2*}(d)\Sigma_{1*}(d)^{-1}$, in which $\sum_{1*}(d) = -E(\partial^2_{\theta(d)}\log p(Y(d)|\mathbf{x}, \theta_*(d)))$ and $\Sigma_{2*}(d) = E(\{\partial_{\boldsymbol{\theta}(d)}\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}_*(d))\}^{\otimes 2})$;
- **d.** For any $\varepsilon_0 > 0$, $\lim_{n\to\infty} P(|K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n) - \mathbf{1}(\Delta_*(d, d') = \mathbf{0})| > \varepsilon_0) = 0$.

Theorem 1 (a) and (b) characterize the asymptotic behavior of $D_{\boldsymbol{\theta}}(d, d'; h_0)$ and $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n)$. Theorem 1 (c) and (d) shows that if the two voxels $d$ and $d'$ have the same true values, then $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n)$ and $\omega(d, d'; h_0)$ converge to 1 and $K_{loc}(\|d - d'\|_2/h_1)$, respectively. However, if the two voxels $d$ and $d'$ substantially differ from each other, then $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n)$ imposes a decreasing weight on the voxel $d'$. As an example, when $K_{st}(u) = \exp(-u)$ and $\lim_{n\to\infty} C_n^{-1}\log(N(\mathcal{D})) = \lim_{n\to\infty} C_n/n = 0$, $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n)$ converges to zero at the rate of $\exp(-C_n^{-1}n)$ when $\boldsymbol{\theta}_*(d) \neq \boldsymbol{\theta}_*(d')$, whereas it converges to 1 at the rate of $\log(N(\mathcal{D}))C_n^{-1}$ otherwise. In the interior of a nonhomogeneous region, $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n)$ automatically puts small weight on the voxels $d'$ with $\boldsymbol{\theta}_*(d) \neq \boldsymbol{\theta}_*(d')$, and thus in the estimation step (ii), the contribution of these voxels $d'$ to the estimation of $\boldsymbol{\theta}_*(d)$ is negligible. Thus, if $\lim_{u\to\infty} K_{st}(u) = 0$ and $\lim_{u\to 0} K_{st}(u) = c$, where $c > 0$ is a fixed scalar, then $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n)$ can efficiently incorporate information from 'good' voxels, while it prevents incorporating information from 'bad' voxels. In contrast, other kernels with $\lim_{u\to\infty} K_{st}(u) > 0$ do not have these features.

For $h > 0$, we can also establish important theoretical results to characterize the attractive behavior of $\hat{\boldsymbol{\theta}}(d, h)$ and $W_\mu(d, h)$ from MARM as follows.

**Theorem 2—**Suppose assumptions (C1)–(C7) in the supplementary report are true. As h > 0, we have the following results for MARM:

- **a.** $\hat{\boldsymbol{\theta}}(d, h)$ converges to $\boldsymbol{\theta}_*(d)$ in probability;
- **b.** $\{\Sigma_{n,2}(\hat{\boldsymbol{\theta}}(d, h))\}^{-1/2}\Sigma_{n,1}(\hat{\boldsymbol{\theta}}(d, h))\{\hat{\boldsymbol{\theta}}(d, h) - \boldsymbol{\theta}_*(d)\} \to^{L} N(0, \mathbf{I}_p)$,

**c.** If $R(\boldsymbol{\theta}*(d)) = \mathbf{b}_0$ is true and $\partial_{\boldsymbol{\theta}(d)} R(\boldsymbol{\theta}*(d))$ is of full rank, then the statistic $W_\mu(d, h)$ is asymptotically distributed as $\chi^2(r)$, a chi-square distribution with r degrees of freedom.

Theorem 2 shows that the MAET procedure has several remarkable features. Theorem 2 (a) ensures that $\hat{\boldsymbol{\theta}}(d, h)$ is a consistent estimate of $\boldsymbol{\theta}*(d)$ for the adaptive weights in (15) for any $h > 0$. Theorem 2 (b) ensures that $\hat{\boldsymbol{\theta}}(d, h)$ is a $\sqrt{n}$ estimate of $\boldsymbol{\theta}*(d)$. Theorem 2 (c) ensures that the Wald test statistic $W_\mu(d, h_s)$ is asymptotically $\chi^2(r)$ distributed under the null hypothesis $R(\boldsymbol{\theta}*(d)) = \mathbf{b}_0$. However, for small sample sizes $n$, it would be better to adjust for sample uncertainty in estimating the covariance matrix of $\hat{\boldsymbol{\theta}}(d, h)$. Following Hotelling's $T^2$ test, we suggest calibrating $W_\mu(d, h)$ with a critical value of $r(n-1)F^{1-\alpha}_{r,n-r}/(n-r)$, where $F^{1-\alpha}_{r,n-r}$ is the upper $\alpha$-percentile of the $F_{r,n-r}$ distribution. That is, we reject $H_0$ if $W_\mu(d, h) \geq r(n-1)F^{1-\alpha}_{r,n-r}/(n-r)$, and do not reject $H_0$ otherwise.

We can characterize the asymptotic behavior of $\hat{\boldsymbol{\theta}}(d, h)$ and $W_\mu(d, h)$ even when $C_n$ is bounded. Our results show the unpleasant behavior of $\hat{\boldsymbol{\theta}}(d, h)$ and $W_\mu(d, h)$ when $h > 0$.

**Corollary 1—**Suppose assumptions (C1)–(C6) in the supplementary report are true, $\lim_{n\to\infty} \log(N(\mathcal{D}))/n = 0$, and $C_n = O(1)$. Then we have the following results:

**a.** $\hat{\boldsymbol{\theta}}(d, h_1)$ converges to $\boldsymbol{\theta}*(d)$ in probability;

**b.** If there is a $d' \in B(d, h_1)/\{d\}$ such that $\boldsymbol{\theta}*(d) = \boldsymbol{\theta}*(d')$, then $\hat{\boldsymbol{\theta}}(d, h_1)$ may not be asymptotically normal and the statistic $W_\mu(d, h_1)$ is not asymptotically distributed as $\chi^2(r)$ even though $R(\boldsymbol{\theta}*(d)) = \mathbf{b}_0$ is true.

Corollary 1 (a) ensures that the PS approach based on a bounded $C_n$ is valid for imaging construction, since $\hat{\boldsymbol{\theta}}(d, h_1)$ is a consistent estimate of $\boldsymbol{\theta}*(d)$. However, Corollary 1 (b) also shows that a bounded $C_n$ can lead to several unpleasant consequences for carrying out statistical inference on $\boldsymbol{\theta}(d)$. Although a bounded $C_n$ has been proposed in the PS approach to smooth the parameter estimates from linear models, we have established here the consistency of $\hat{\boldsymbol{\theta}}(d, h)$ as an estimate of $\boldsymbol{\theta}*(d)$ under a general setup. Moreover, if there is a voxel $d' \in B(d, h_1)/\{d\}$ such that $\boldsymbol{\theta}*(d) = \boldsymbol{\theta}*(d')$, Corollary 1 (b) shows that $\hat{\boldsymbol{\theta}}(d, h_1)$ is not asymptotically normal and the Wald test statistic $W_\mu(d, h_1)$ is not asymptotically $\chi^2(r)$ distributed under the null hypothesis $R(\boldsymbol{\theta}*(d)) = \mathbf{b}_0$. Thus, we cannot directly calibrate $W_\mu(d, h_1)$ using the critical values of $\chi^2(r)$.

Finally, we focus on a multiscale adaptive linear model. Assume that $Y_i(d) = \mathbf{x}_i^T \beta(d) + \varepsilon_i(d)$, where $\varepsilon_i(d) \sim N(0, \tau(d)^{-1})$. Let $\tilde{\omega}_\tau(d, d'; h) = \tau(d')\omega(d, d'; h)/\Sigma_{d' \in B(d,h)} \tau(d')\omega(d, d'; h)$, we have

$$\widehat{\beta}(d,h)=\Big(\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2}\Big)^{-1}\sum_{i=1}^{n}\mathbf{x}_i Y_i(d;\tilde{\omega}_\tau,h) \text{ and } \mathrm{Cov}(\widehat{\beta}(d,h)) \approx \Big(\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2}\Big)^{-1}\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2}\widehat{\varepsilon}_i(d;\tilde{\omega}_\tau,h)^2\Big(\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2}\Big)^{-1}, \tag{19}$$

where $Y_i(d; \tilde{\boldsymbol{\omega}}_\tau, h) = \Sigma_{d' \in B(d,h)} \tilde{\omega}_\tau(d, d'; h)\, Y_i(d')$ and $\widehat{\varepsilon}_i(d;\tilde{\omega}_\tau,h)=\sum_{d' \in B(d,h)} \tilde{\omega}_\tau(d,d';h)[Y_i(d') - \mathbf{x}_i^T\widehat{\beta}(d',h)]$. Although Tabelow *et al*. (2006) have obtained the same $\hat{\boldsymbol{\beta}}(d, h)$ as in (19), the MARM developed here has several advantages. We will show below that $\hat{\boldsymbol{\beta}}(d, h)$ based on the adaptive weights in the PS approach may not be asymptotically normal. The covariance estimate of $\hat{\boldsymbol{\beta}}(d, h)$ in (19) has a simple form. We

obtain the following results for the multiscale adaptive linear model. For simplicity, we assume that all $\tau(d)$ are known.

### Theorem 3

**a.** If assumptions (C1), (C2), (C6) and (C7) in the supplementary report are true, $E(\|\mathbf{x}\|_2^2)<\infty$ and $E(\max_{d\in\mathcal{D}}|\varepsilon(d)|^2\times\|\mathbf{x}\|_2^2)<\infty$, then $\sqrt{n}\{\widehat{\beta}(d,h)-\beta_*\}$ is asymptotically equivalent to

$$A_1(d;h)=\sum_{d'\in B(d,h)}C(d,d';h)\tau(d')E(\mathbf{x}^{\otimes 2})^{-1}n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i(d')/\{\sum_{d'\in B(d,h)}C(d,d';h)\tau(d')\}, \tag{20}$$

where $C(d, d'; h) = \mathbf{1}(\Delta_*(d, d') = \mathbf{0})K_{loc}(\|d- d'\|_2/h)$. The $A_1(d; h)$ converges in distribution to

$$\sum_{d'\in B(d,h)}C(d,d';h)\tau(d')E(\mathbf{x}^{\otimes 2})^{-1/2}Z(d')/\{\sum_{d'\in B(d,h)}C(d,d';h)\tau(d')\}, \tag{21}$$

where $\{Z(d'): d' \in B(d, h)\}$ is a Gaussian vector with mean zero and covariance structure $\mathrm{Cov}(Z(d)) = \tau(d)^{-1}\mathbf{I}_{p1}$ and $\mathrm{Cov}(Z(d), Z(d')) = E(\varepsilon_1(d)\varepsilon_1(d'))\mathbf{I}_{p1}$.

**b.** If assumptions (C1), (C2) and (C6) in the supplementary report are true, $C_n = O(1)$ and $\lim_{n\to\infty}\log(N(\mathcal{D}))/n = 0$, then $\sqrt{n}\{\widehat{\beta}(d,h_1)-\beta_*\}$ is asymptotically equivalent to

$$A_2(d;h_1)=\frac{\sum_{d'\in B(d,h_1)}C(d,d';h_1)K_{st}(\mathcal{E}_n(d,d'))\tau(d')E(\mathbf{x}^{\otimes 2})^{-1}n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i(d')}{\sum_{d'\in B(d,h_1)}C(d,d';h_1)K_{st}(\mathcal{E}_n(d,d'))\tau(d')},$$

where $\mathcal{E}_n(d,d')=\tau(d)tr([E(\mathbf{x}^{\otimes 2})^{-1/2}n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\{\varepsilon_i(d)-\varepsilon_i(d')\}]^{\otimes 2})$. As $n\to\infty$ $A_2(d; h_1)$ converges in distribution to a random vector given by

$$\frac{\sum_{d'\in B(d,h)}C(d,d';h_1)K_{st}(\tau(d)tr(\{Z(d)-Z(d')\}^{\otimes 2}))\tau(d')E(\mathbf{x}^{\otimes 2})^{-1/2}Z(d')}{\sum_{d'\in B(d,h)}C(d,d';h_1)K_{st}(\tau(d)tr(\{Z(d)-Z(d')\}^{\otimes 2}))\tau(d')}.$$

Theorem 3 gives a theoretical justification of the multiscale adaptive linear model. Theorem 3 (a) and (b) formally characterize the key differences between a bounded and unbounded $C_n$ in the linear model. Theorem 3 (a) shows that for certain unbounded $C_n$, the asymptotic distributions of $\hat{\boldsymbol{\beta}}(d, h)$ are always normally distributed. For a bounded $C_n$, however, Theorem 3 (b) only gives the asymptotic distribution of $\hat{\boldsymbol{\beta}}(d, h_1)$, which may not be normally distributed when there is a voxel $d' \in B(d, h_1)$ whose data are close to those of the voxel $d$.

## 3. Simulation Studies

We conducted three sets of Monte Carlo simulations to examine the finite sample performance of $\hat{\boldsymbol{\beta}}(d, h)$ and $W_\mu(d, h)$ with respect to different scales $h$ and compare MARM with the voxel-wise method. For the sake of space, we only present some results based on a $64 \times 64$ phantom image with four known effect regions and put additional simulation results in the supplementary document.

We simulated data at all $m = 4096$ pixels on the 64×64 phantom image for $n$ subjects. At a given pixel $d$ in $\mathcal{D}$, $Y_i(d)$ was simulated according to $Y_i(d)=\mathbf{x}_i^T\beta(d)+\varepsilon_i(d)$ for $i = 1, \cdots, n$, where $\boldsymbol{\beta}(d) = (\beta_1(d), \beta_2(d), \beta_3(d))^T$ and $\mathbf{x}_i = (1, x_{i2}, x_{i3})^T$. Errors $\varepsilon_i(d)$ were first independently generated from $N(0, 1)$ and $\chi^2(3) - 3$, respectively, and then they were smoothed by using heat kernel smoothing with 4 iterations, which gave an effective smoothness of about 2 pixels (Chung *et al.*, 2005). The $\chi^2(3) - 3$ distribution is a very skewed distribution. We set $n = 60$ and $n = 80$. We generated $x_{i2}$ independently from a Bernoulli distribution with probability of success being 0.5, and generated $x_{i3}$ independently from the uniform distribution on [1, 2]. The $x_{i2}$ and $x_{i3}$ were chosen to represent group identity and scaled age, respectively. Furthermore, we set $\beta_1(d) = \beta_3(d) = 0$ across all pixels $d$. For $\beta_2(d)$, we divided the 64×64 phantom image into five different regions of interest (ROIs) with different shapes and then varied $\beta_2(d)$ as 0, 0.2, 0.4, 0.6 and 0.8, respectively, across these five ROIs. Different $\beta_2(d)$ values, which represent different signal-to-noise ratios, were chosen to examine the performance of our method at different signal-to-noise ratios and also to test whether MARM can perform well for different shapes. The true $\beta_2(d)$ was displayed for all ROIs with black, blue, red, yellow, and white colors representing $\beta_2(d)$=0, 0.2, 0.4, 0.6, and 0.8, respectively (Fig. 2(k)).

We fitted the linear model $Y_i(d)=\mathbf{x}_i^T\beta(d)+\varepsilon_i(d)$, where $\varepsilon_i(d) \sim N(0, \tau(d)^{-1})$, and then applied the MAET procedure described in Example 4 to calculate adaptive parameter estimates across all pixels at 11 different scales. Next, for $\beta_2(d)$, we calculated the bias, the empirical standard error (RMS), the mean of the standard error estimates (SD), the ratio of RMS over SD (RE), and the achievable variance reduction (VR), defined as $\mathrm{Var}(\hat{\beta}_2(d, h_s))/\mathrm{Var}(\hat{\beta}_2(d, h_0))$, at each pixel of all five ROIs based on the results obtained from the 1,000 simulated data sets. For the sake of space, we only presented the results of $\hat{\beta}_2(d, h_0)$ and $\hat{\beta}_2(d, h_{10})$ obtained from $N(0, 1)$ distributed data with $n = 60$ in Fig. 2. We also calculated the average bias, RMS, SD, RE, and maximum VR (MVR) in each of the five ROIs and presented them in Table 1. The biases are slightly increased from $h_0$ to $h_{10}$ (Fig. 2(b) and (g) and Table 1), whereas RMS and SD at $h_5$ and $h_{10}$ are much smaller than those at $h_0$ (Fig. 2(c), (d), (h), (i) and 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 and Fig. 2(e) and (j)). Moreover, the SDs in these pixels of ROIs with $\beta_2(d) > 0$ are larger than the SDs in those pixels of ROI with $\beta_2(d) = 0$ (Figs. 2(i)), because the interior of ROI with $\beta_2(d) = 0$ contains more pixels (Fig. 2(k)). The biases, SDs, RMSs, and MVRs of $\beta_2(d)$ are smaller in the normally distributed data than in the chi-square distributed data (Table 1), because the signal-to-noise ratios (SNRs) in the normally distributed data are 2.45 times bigger than SNRs in the chi-square distributed data. Increasing the sample size and SNR decreases the bias, RMS, SD, and MVR of the parameter estimates (Table 1).

We then tested the hypotheses $H_0$: $\beta_2(d) = 0$ and $H_1$: $\beta_2(d) \neq 0$ across all pixels to assess both Type I and II error rates at the pixel level. We applied the same MAET procedure and computed the $p$-values of $W_\mu(d, h)$ at each scale. The 1,000 replications were used to calculate the estimates and standard errors of rejection rates at $\alpha = 5\%$ significance level. For $W_\mu(d, h)$, the Type I error rates in ROI with $\beta_2(d) = 0$ were relatively accurate for all scales, while the statistical power for rejecting the null hypothesis in ROIs with $\beta_2(d) \neq 0$ was significantly increased with radius $h$ and SNR (Table 2).

## 4. Real Data Analysis

Understanding white matter development in the human brain *in vivo* is critical to the understanding of the functional formation of the central nervous system. An important feature of diffusion tensor imaging (DTI) is its capability to reveal the white matter

maturation process in the human brain using a set of water diffusion related parameters, such as fractional anisotropy (FA) and radial (RD) diffusivity. For instance, FA represents the inhomogeneous extent of local barriers to water diffusion and has been widely used to investigate early brain development from identifying transient brain structures such as ganglionic eminence and cortical subplate as well as estimating the correlation of white matter maturation with functional development measures such as IQ and working memory.

We considered 38 subjects from the neonatal project on early brain development led by Dr. Gilmore at the University of North Carolina at Chapel Hill. For each subject, diffusion-weighted images were acquired at 2 weeks, year 1 and year 2. Diffusion tensor acquisition scheme includes 18 repeated measures of six non-collinear directions, (1,0,1), (−1,0,1), (0,1,1), (0,1,−1), (1,1,0), and (−1,1,0) at a b-value of 1000 s/mm2 and a b=0 reference scan. Forty-six contiguous slices with a slice thickness of 2 mm covered a field of view (FOV) of 256×256 $mm^2$ with an isotropic voxel size of $2 \times 2 \times 2$ $mm^3$. High resolution T1 weighted (T1W) images were acquired using a 3D MP-RAGE sequence. Then, a weighted least squares estimation method was used to construct the diffusion tensors (Basser *et al.*, 1994; Zhu *et al.*, 2007b). All DT images (38 subjects, 3 time points each) were registered to a randomly selected brain DTI of a 2-year-old subject using a tensor image morphing for elastic registration (TIMER) (Yap *et al.*, 2009).

Fractional anisotropy (FA) calculated from DTIs is widely used as a measurement to assess directional organization of the brain which is greatly influenced by the magnitude and orientation of white matter tracts. We used FA images to identify the spatial patterns of white matter maturation, and then considered a multivariate linear model

$Y_{ij}(d)=\beta_1(d)+t_{ij}\beta_2(d)+t_{ij}^2\beta_3(d)+\varepsilon_{ij}(d)$ for $i = 1, \cdots, 38$ and $j = 1, 2, 3$, at each voxel of the template, where $t_{ij}$ denotes the $j$-th scan time for the $i$-th subject, $\varepsilon_i(d) = (\varepsilon_{i1}(d), \varepsilon_{i2}(d), \varepsilon_{i3}(d))^T \sim N(\mathbf{0}, \Sigma(d))$, and $\Sigma(d)$ is a 3×3 unstructured covariance matrix. The MAET procedure described in Example 4 with $c_h = 1.15$ and $S = 10$ was used to carry out statistical analysis. We tested $H_0$: $\beta_2(d) = \beta_3(d) = 0$ for age dependent effects across all voxels $d$ and calculated the corrected $p$-values using the Bonferroni correction with overall significance level of 1%. As $s$ increases from 0 to 10, MARM shows a clear advantage in detecting more significant and smoothed significant areas as well as preserving the edges of gray matter, white matter, and cerebrospinal fluid areas (Fig. 3(a)–(d) and (h)). We also smoothed FA imaging data using an isotropic Gaussian kernel with FWHM 6mm and then analyzed the data using the voxel-wise approach. The results based on the smoothed FA images show the obvious oversmoothing in CSF and the gray matter areas, such as the ventricle (Fig. 3(a)–(c)). Furthermore, we identified a voxel in the red circle in the ventricle, whose location is near the boundary of the white matter and CSF (see red circle in Fig. 3(a)). Its corrected $p$ values of $W_\mu(d, h_0)$ and $W_\mu(d, h_{10})$ are much higher than 0.01. Inspecting raw FA values in the red voxel of Figure 3(a) does not reveal any growth patterns, which agrees with the fact that the ventricle contains CSF in the brain (Fig. 3(d)). However, after being smoothed with the Gaussian kernel, smoothed FA values gradually increase with age (Fig. 3(h)). This indicates that the data in the red voxel was oversmoothed due to its neighboring voxels containing white matter.

The parameters $\beta_1(d)$, $\beta_2(d)$, and $\beta_3(d)$ represent the FA value at birth (age = 0) and the speed and acceleration of the change of FA, respectively (Figs. 3(e)–(g)). Major white matter structures are already presented in FA at birth (Fig. 3(e)). Within the central brain region, results show different developing patterns for the genu, splenium and body of corpus callosum, internal and external capsules (Figs. 3(i)–(l)). Comparing FA values, while genu and splenium have a similar FA value at birth, results indicate that genu's FA gradually increases to a higher value than splenium's over time. The corpus callosum body has a slightly lower FA compared to the internal capsule at birth, but gradually surpasses the

internal capsule. The external capsule, having the lowest FA value among these white matter regions at birth, demonstrates a slow linear-like changing pattern.

## 5. Discussion

This article studies the idea of using a multiscale adaptive regression model for the spatial and adaptive analysis of neuroimaging data. MARM integrates the PS approach with the voxel-wise method for neuroimaging data from multiple subjects. There are three features in MARM: being spatial, being hierarchical and being adaptive. MARM builds a sphere with a given radius at all voxels, and then uses these consecutively overlapping spheres to capture local and global spatial dependence among different voxels. Thus, MARM explicitly utilizes the spatial information to carry out statistical inference. MARM also builds hierarchically nested spheres by increasing the radius of a spherical neighborhood around each voxel and utilizes information in each of the nested spheres across all voxels. Finally, MARM combines all observations with adaptive weights in the voxels within the sphere of the current voxel to adaptively calculate parameter estimates and test statistics. Without imposing any spatial correlation patterns, we have derived the asymptotic properties of the parameter estimates and test statistics for MARM when the logarithm of the number of voxels is relatively small compared with the number of subjects. We also investigated the issue of selecting the appropriate values of various parameters in MAET.

Many issues still merit further research. The three key features of MARM can be easily adapted to more complex data structures (e.g., longitudinal, twin and family) and other parametric and semiparametric models. For instance, for longitudinal neuroimaging data, we can develop a multiscale adaptive method for generalized estimating equations. It is also feasible to consider statistical models with nonparametric components. More research is needed for optimizing the choices of parameters in MAET and regularity assumptions. For instance, by assuming spatial smoothness in the neuroimaging data, the assumption $\log(N(\mathcal{D})) \ll C_n \ll n$ can be weakened. Another interesting issue that warrants future investigation is the development of methods which determine multiscale neighborhoods adaptive to the pattern of imaging data at each voxel. An important issue is that the voxel-wise approach and MARM are also based on the perfect registration assumption, that is demonstrably false. We may need to integrate the registration method, smoothing method, and voxel-wise approach into a unified framework so that we can appropriately account for registration errors in the statistical analysis.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank the Editor, an Associate Editor and two referees for valuable suggestions, which helped to improve our presentation greatly. Thanks to Ms. Diana Lam and Martha Skup for her invaluable editorial assistance. This work was supported in part by NIH grants UL1-RR025747-01, P01CA142538-01, GM70335, CA74015, MH086633, AG033387, MH064065, HD053000, MH070890, and R01NS055754.

## References

Ashburner J, Friston KJ. Voxel-based morphometry: the methods. NeuroImage. 2000; 11:805–821. [PubMed: 10860804]

Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. Chapman Hall; Boca Raton: 2004.

Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. Biophysical Journal. 1994; 66:259–267. [PubMed: 8130344]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Ser B. 1995; 57:289–300.

Besag JE. On the statistical analysis of dirty pictures (with discussion). Journal of the Royal Statistical Society, Ser B. 1986; 48:259–302.

Bowman FD. Spatio-temporal models for region of interest analyses of functional mappping experiments. Journal of American Statistical Association. 2007; 102:442–453.

Chung MK, Robbins S, Dalton KM, Davidson RJ, Alexander AL, Evans AC. Cortical thickness analysis in autism via heat kernel smoothing. NeuroImage. 2005; 25:1256–1265. [PubMed: 15850743]

Friston, KJ. Statistical Parametric Mapping: the Analysis of Functional Brain Images. Academic Press; London: 2007.

Hecke WV, Sijbers J, Backer SD, Poot D, Parizel PM, Leemans A. On the construction of a ground truth framework for evaluating voxel-based diffusion tensor MRI analysis methods. NeuroImage. 2009; 46:692–707. [PubMed: 19268708]

Huettel, SA.; Song, AW.; McCarthy, G. Functional Magnetic Resonance Imaging. Sinauer Associates, Inc; 2004.

Jones DK, Symms DK, Cercignani M, Howard RJ. The effect of filter size on VBM analyses of DT-MRI data. NeuroImage. 2005; 26:546–554. [PubMed: 15907311]

Lazar, N. The Statistical Analysis of Functional MRI Data. New York: Springer; 2008.

McCullagh, P.; Nelder, JA. Generalized Linear Models. 2. London: Chapman and Hall; 1989.

Nichols T, Hayasaka S. Controlling the family-wise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research. 2003; 12:419–446. [PubMed: 14599004]

Polzehl J, Spokoiny VG. Adaptive weights smoothing with applications to image restoration. J R Statist Soc B. 2000; 62:335–354.

Polzehl J, Spokoiny VG. Propagation-separation approach for local likelihood estimation. Probab Theory Relat Fields. 2006; 135:335–362.

Polzehl J, Tabelow K. fmri: A package for analyzing fmri data. R News. 2007; 7:13–17.

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

Qiu P. Jump surface estimation, edge detection, and image restoration. Journal of American Statistical Association. 2007; 102:745–756.

Tabelow K, Piech V, Polzehl J, Voss HU. High-resolution fMRI: overcoming the signal-to-noise problem. J Neurosci Meth. 2008a; 178:357–365.

Tabelow K, Polzehl J, Voss HU, Spokoiny V. Analyzing fMRI experiments with structural adaptive smoothing procedures. NeuroImage. 2006; 33:55–62. [PubMed: 16891126]

Tabelow K, Polzehl J, Spokoiny V, Voss HU. Diffusion tensor imaging: structural adaptive smoothing. NeuroImage. 2008b; 39:1763–1773. [PubMed: 18060811]

Tabelow K, Polzehl J, Ulug AM, Dyke JP, Watts R, Heier LA, Voss HU. Accurate localization of functional brain activity using structure adaptive smoothing. IEEE Trans Med Imaging. 2008c; 27:531–537. [PubMed: 18390349]

Thompson PM, Toga AW. A framework for computational anatomy. Computing and Visualization in Science. 2002; 5:13–34.

Worsley KJ. Detecting activation in fMRI data. Statistical Methods in Medical Research. 2003; 12:401–418. [PubMed: 14599003]

Worsley KJ, Taylor JE, Tomaiuolo F, Lerch J. Unified univariate and multivariate random field theory. NeuroImage. 2004; 23:189–195.

Yap PT, Wu GR, Zhu HT, Lin WL, Shen DG. TIMER: tensor image morphing for elastic registration. NeuroImage. 2009; 47:549–563. [PubMed: 19398022]

Yue Y, Loh JM, Lindquist MA. Adaptive spatial smoothing of fMRI images. Statistics and its Interface. 2010; 3:3–14.

Zhu HT, Gu MG, Peterson BG. Maximum likelihood from spatial random effects models via the stochastic approximation expectation maximization algorithm. Statistics and Computing. 2007a; 15:163–177.

Zhu HT, Zhang HP, Ibrahim JG, Peterson BS. Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance image data (with discussion). Journal of the American Statistical Association. 2007b; 102:1085–1102.

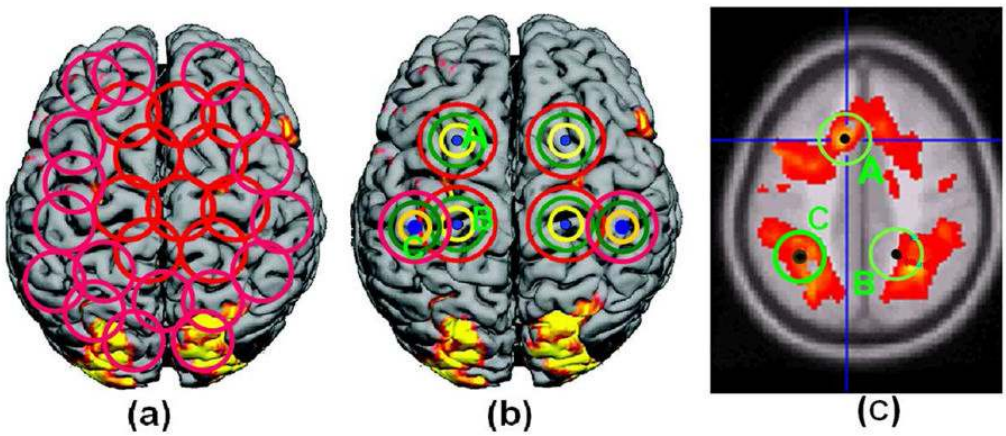


**Fig. 1.**

Illustration of the key features in the multiscale adaptive regression model. For a relatively large radius $r_0$, panel (a) shows the overlapping spherical neighborhoods $B(d, r_0)$ of multiple points (or voxels) $d$ on the cortical surface. Panel (b) shows the spherical neighborhoods with four different bandwidths $h$ of the six selected points $d$ on the cortical surface. Panel (c) shows the spherical neighborhoods $B(d, r_0)$ of three selected voxels in a 3D volume, in which voxels A and C are inside the activated regions, whereas voxel B is on the boundary of an activated region.

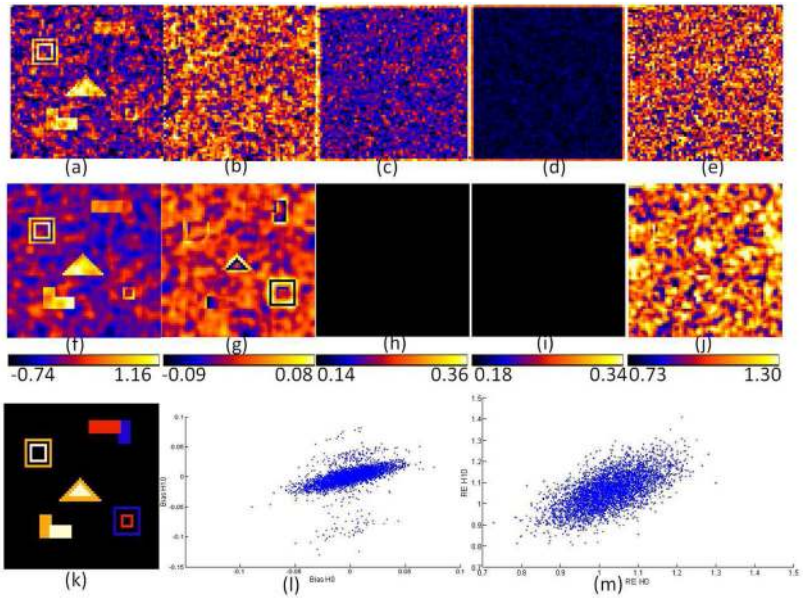


**Fig. 2.**

Results from a simulation study of comparing the voxel-wise method and MARM based on 1,000 $N(0, 1)$ distributed data with $n = 60$. Panel (k) is the ground truth image of five ROIs with black, blue, red, yellow, and white color representing $\beta_2(d)$=0, 0.2, 0.4, 0.6, and 0.8, respectively. The first row contains the results from the voxel-wise method: (a) a selected image of $\hat{\beta}_2(d, h_0)$ obtained from a simulated data set; (b) bias image of $\hat{\beta}_2(d, h_0)$; (c) RMS image of $\hat{\beta}_2(d, h_0)$; (d) SD image of $\hat{\beta}_2(d, h_0)$; and (e) RE image of $\hat{\beta}_2(d, h_0)$. The second row contains the results obtained from MAET as $S = 10$ and $c_h = 1.1$: (f) a selected image of $\hat{\beta}_2(d, h_{10})$ obtained from a simulated data set; (g) bias image of $\hat{\beta}_2(d, h_{10})$; (h) RMS image of $\hat{\beta}_2(d, h_{10})$; (i) SD image of $\hat{\beta}_2(d, h_{10})$; and (j) RE image of $\hat{\beta}_2(d, h_{10})$. Panels (l) and (m) are the scatter plots of biases and REs of $\hat{\beta}_2(d, h_0)$ versus $\hat{\beta}_2(d, h_{10})$, respectively.

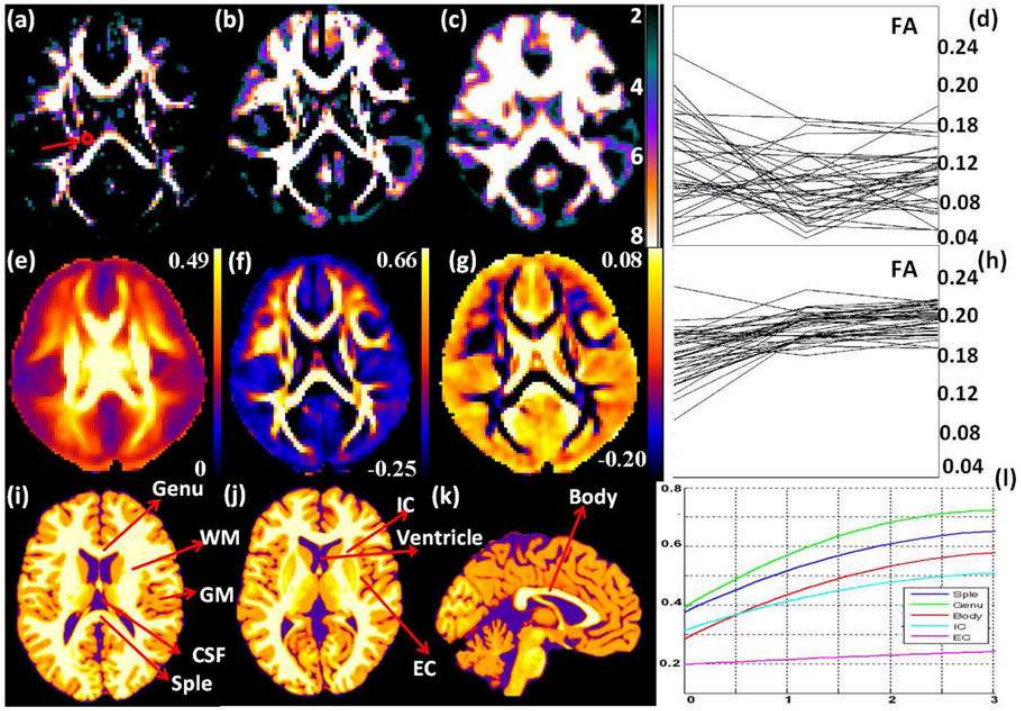


**Fig. 3.**

Results from the neonatal project on brain development. Panel (a): the Bonferroni corrected $-\log_{10}(p)$ values of $W_\mu(d, h_0)$ from a selected slice and a selected voxel in the red circle in the ventricle; panel (b): the Bonferroni corrected $-\log_{10}(p)$ values of $W_\mu(d, h_{10})$ from the same selected slice; panel (c): the Bonferroni corrected $-\log_{10}(p)$ values of the Wald test statistics obtained from the Gaussian kernel smoothed FA images for the same selected slice; panel (d): longitudinal trajectories of unsmoothed FA values in the red voxel identified in panel (a); panel (h): longitudinal trajectories of the Gaussian kernel smoothed FA values in the red voxel identified in panel (a); panels (e), (f), and (g): estimated $\hat{\beta}_1(d, h_{10})$, $\hat{\beta}_2(d, h_{10})$, and $\hat{\beta}_3(d, h_{10})$ for the same selected slice; panels (i), (j), and (k): anatomical images with eight labeled regions of interest including the genu, splenium (Sple), internal capsule (IC), external capsule (EC), ventricle, grey matter (GM), white matter (WM), cerebrospinal fluid (CSF), and corpus callosum body (Body); panel (l): the growth patterns from the ROIs located in the splenium (Sple), genu (Genu) and body (Body) of corpus callosum, internal capsule (IC), and external capsule (EC) for FA.

**Table 1**

Average Bias ($\times 10^{-3}$), RMS, SD, RE, and MVR of $\boldsymbol{\beta}_2(d)$ parameters in the five ROIs at 3 different scales ($h_0$, $h_5$, $h_{10}$), 2 different distributions ($N(0, 1)$ and $\chi^2(3) - 3$ distributions), 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; MVR denotes the maximum achievable variance reduction. For each case, 1,000 simulated data sets were used.

| | $\chi^2(3) - 3$ | | | | | | $N(0, 1)$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n = 60$ | | | $n = 80$ | | | $n = 60$ | | | $n = 80$ | | |
| | $h_0$ | $h_5$ | $h_{10}$ | $h_0$ | $h_5$ | $h_{10}$ | $h_0$ | $h_5$ | $h_{10}$ | $h_0$ | $h_5$ | $h_{10}$ |
| | $\beta_2(d) = 0.0$ | | | | | | $\beta_2(d) = 0.0$ | | | | | |
| BIAS | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| RMS | 0.48 | 0.35 | 0.26 | 0.41 | 0.31 | 0.22 | 0.20 | 0.15 | 0.11 | 0.17 | 0.13 | 0.09 |
| SD | 0.47 | 0.34 | 0.24 | 0.41 | 0.30 | 0.21 | 0.19 | 0.14 | 0.10 | 0.17 | 0.12 | 0.09 |
| RE | 1.03 | 1.05 | 1.06 | 1.02 | 1.03 | 1.04 | 1.03 | 1.05 | 1.06 | 1.02 | 1.03 | 1.04 |
| VMR | 1.00 | 0.59 | 0.44 | 1.00 | 0.61 | 0.46 | 1.0 | 0.63 | 0.46 | 1.0 | 0.64 | 0.47 |
| | $\beta_2(d) = 0.2$ | | | | | | $\beta_2(d) = 0.2$ | | | | | |
| BIAS | 0.00 | −0.03 | −0.07 | 0.01 | −0.02 | −0.06 | 0.00 | −0.03 | −0.05 | 0.00 | −0.02 | −0.05 |
| RMS | 0.46 | 0.34 | 0.24 | 0.39 | 0.29 | 0.21 | 0.19 | 0.14 | 0.11 | 0.16 | 0.12 | 0.09 |
| SD | 0.46 | 0.33 | 0.24 | 0.40 | 0.29 | 0.21 | 0.19 | 0.14 | 0.10 | 0.16 | 0.12 | 0.09 |
| RE | 1.01 | 1.01 | 1.01 | 0.99 | 1.00 | 1.01 | 1.02 | 1.04 | 1.06 | 1.02 | 1.02 | 1.03 |
| VMR | 1.00 | 0.70 | 0.50 | 1.00 | 0.71 | 0.51 | 1.00 | 0.72 | 0.52 | 1.00 | 0.73 | 0.52 |
| | $\beta_2(d) = 0.4$ | | | | | | $\beta_2(d) = 0.4$ | | | | | |
| BIAS | −0.01 | −0.05 | −0.09 | 0.01 | −0.02 | −0.06 | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 |
| RMS | 0.46 | 0.34 | 0.25 | 0.40 | 0.30 | 0.22 | 0.19 | 0.15 | 0.12 | 0.16 | 0.13 | 0.10 |
| SD | 0.46 | 0.33 | 0.24 | 0.40 | 0.29 | 0.21 | 0.19 | 0.14 | 0.11 | 0.16 | 0.12 | 0.09 |
| RE | 1.01 | 1.02 | 1.03 | 1.01 | 1.02 | 1.03 | 1.03 | 1.05 | 1.07 | 1.00 | 1.01 | 1.02 |
| VMR | 1.00 | 0.70 | 0.50 | 1.00 | 0.70 | 0.51 | 1.00 | 0.71 | 0.52 | 1.00 | 0.72 | 0.52 |
| | $\beta_2(d) = 0.6$ | | | | | | $\beta_2(d) = 0.6$ | | | | | |
| BIAS | 0.00 | −0.05 | −0.09 | 0.00 | −0.04 | −0.07 | 0.00 | 0.01 | 0.02 | 0.00 | 0.00 | 0.01 |
| RMS | 0.46 | 0.35 | 0.26 | 0.40 | 0.30 | 0.23 | 0.19 | 0.15 | 0.12 | 0.16 | 0.13 | 0.10 |

| | $\chi^2(3) - 3$ | | | | | | $N(0, 1)$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n = 60$ | | | $n = 80$ | | | $n = 60$ | | | $n = 80$ | | |
| | $h_0$ | $h_5$ | $h_{10}$ | $h_0$ | $h_5$ | $h_{10}$ | $h_0$ | $h_5$ | $h_{10}$ | $h_0$ | $h_5$ | $h_{10}$ |
| SD | 0.46 | 0.34 | 0.25 | 0.40 | 0.30 | 0.22 | 0.19 | 0.14 | 0.11 | 0.16 | 0.13 | 0.10 |
| RE | 1.01 | 1.03 | 1.04 | 1.01 | 1.02 | 1.03 | 1.02 | 1.04 | 1.06 | 1.01 | 1.03 | 1.04 |
| VMR | 1.00 | 0.70 | 0.50 | 1.00 | 0.71 | 0.52 | 1.00 | 0.71 | 0.52 | 1.00 | 0.72 | 0.52 |
| | $\beta_2(d) = 0.8$ | | | | | | $\beta_2(d) = 0.8$ | | | | | |
| BIAS | 0.00 | −0.04 | −0.06 | 0.00 | −0.02 | −0.05 | 0.00 | −0.01 | −0.02 | 0.00 | 0.00 | −0.01 |
| RMS | 0.47 | 0.35 | 0.26 | 0.40 | 0.30 | 0.23 | 0.19 | 0.15 | 0.11 | 0.17 | 0.13 | 0.10 |
| SD | 0.46 | 0.34 | 0.25 | 0.40 | 0.30 | 0.22 | 0.19 | 0.14 | 0.11 | 0.16 | 0.12 | 0.09 |
| RE | 1.02 | 1.03 | 1.04 | 1.01 | 1.02 | 1.03 | 1.02 | 1.04 | 1.05 | 1.03 | 1.05 | 1.06 |
| VMR | 1.00 | 0.71 | 0.51 | 1.00 | 0.71 | 0.51 | 1.00 | 0.71 | 0.51 | 1.00 | 0.73 | 0.52 |



*J R Stat Soc Series B Stat Methodol*. Author manuscript; available in PMC 2011 December 1.

**Table 2**

Simulation study for $W_\mu(d, h)$: estimates (ES) and standard errors (SE) of rejection rates for pixels inside the five ROIs were reported at 2 different scales ($h_0$, $h_{10}$), 2 different distributions ($N$ (0, 1) and $\chi^2(3)-3$), and 2 different sample sizes ($n = 60, 80$) at $\alpha = 5\%$. For each case, 1,000 simulated data sets were used.

| | | $N(0, 1)$ | | | | $\chi^2(3) - 3$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $n = 60$ | | $n = 80$ | | $n = 60$ | | $n = 80$ | |
| $\beta_2(d)$ | $h_s$ | ES | SE | ES | SE | ES | SE | ES | SE |
| 0.2 | $h_0$ | 0.20 | 0.066 | 0.24 | 0.070 | 0.08 | 0.038 | 0.08 | 0.037 |
| | $h_{10}$ | 0.30 | 0.126 | 0.38 | 0.121 | 0.10 | 0.069 | 0.18 | 0.081 |
| 0.4 | $h_0$ | 0.56 | 0.090 | 0.67 | 0.079 | 0.15 | 0.065 | 0.18 | 0.070 |
| | $h_{10}$ | 0.93 | 0.051 | 0.98 | 0.030 | 0.26 | 0.129 | 0.35 | 0.159 |
| 0.6 | $h_0$ | 0.88 | 0.039 | 0.95 | 0.024 | 0.27 | 0.057 | 0.33 | 0.050 |
| | $h_{10}$ | 1.00 | 0.004 | 1.00 | 0.004 | 0.51 | 0.091 | 0.63 | 0.083 |
| 0.8 | $h_0$ | 0.99 | 0.015 | 1.00 | 0.005 | 0.43 | 0.080 | 0.52 | 0.080 |
| | $h_{10}$ | 0.99 | 0.010 | 0.99 | 0.011 | 0.78 | 0.099 | 0.90 | 0.006 |
| 0.0 | $h_0$ | 0.07 | 0.006 | 0.07 | 0.006 | 0.06 | 0.007 | 0.07 | 0.006 |
| | $h_{10}$ | 0.08 | 0.011 | 0.07 | 0.011 | 0.07 | 0.012 | 0.08 | 0.012 |