

# MARM: Multiscale Adaptive Regression Models for Neuroimaging Data

Hongtu Zhu<sup>1,3</sup>, Yimei Li<sup>1</sup>, Joseph G. Ibrahim<sup>1</sup>, Weili Lin<sup>2,3</sup>,  
and Dinggang Shen<sup>2,3</sup>

<sup>1</sup> Departments of Biostatistics, <sup>2</sup> Radiology, <sup>3</sup> Biomedical Research Imaging Center,  
University of North Carolina at Chapel Hill

**Abstract.** We develop a novel statistical model, called *multiscale adaptive regression model (MARM)*, for spatial and adaptive analysis of neuroimaging data. The primary motivation and application of the proposed methodology is statistical analysis of imaging data on the two-dimensional (2D) surface or in the 3D volume for various neuroimaging studies. The existing voxel-wise approach has several major limitations for the analyses of imaging data, underscoring the great need for methodological development. The voxel-wise approach essentially treats all voxels as independent units, whereas neuroimaging data are spatially correlated in nature and spatially contiguous regions of activation with rather sharp edges are usually expected. The initial smoothing step before the voxel-wise approach often blurs the image data near the edges of activated regions and thus it can dramatically increase the numbers of false positives and false negatives. The MARM, which is developed for addressing these limitations, has three key features in the analysis of imaging data: being spatial, being hierarchical, and being adaptive. The MARM builds a small sphere at each location (called voxel) and use these consecutively connected spheres across all voxels to capture spatial dependence among imaging observations. Then, the MARM builds hierarchically nested spheres by increasing the radius of a spherical neighborhood around each voxel and combine all the data in a given radius of each voxel with appropriate weights to adaptively calculate parameter estimates and test statistics. Theoretically, we first establish that the MARM outperforms classical voxel-wise approach. Simulation studies are used to demonstrate the methodology and examine the finite sample performance of the MARM. We apply our methods to the detection of spatial patterns of brain atrophy in a neuroimaging study of Alzheimers disease. Our simulation studies with known ground truth confirm that the MARM significantly outperforms the voxel-wise methods.

## 1 Introduction

Anatomical and functional magnetic resonance imaging (MRI) are powerful tools for understanding the neural development of neuropsychiatric disorders, substance use disorders, and normal brains. Specifically, anatomical MRI has

been widely used to segment the cortical and subcortical structures (e.g., hippocampus) of the human brain *in vivo* and to generate various morphological measures of their morphology for understanding neuroanatomical differences in brain structure across different populations [1]. Functional MRI (fMRI) has been widely used to understand functional integration of different brain regions in response to specific stimuli and behavioral tasks and detecting the association between brain function and covariates of interest, such as diagnosis, behavioral tasks, severity of disease, age, or IQ [2,3,4].

Much effort has been devoted to developing voxel-wise methods for analyzing various imaging measures including cortical thickness using numerical simulations and theoretical reasoning. The voxel-wise methods for analyzing imaging data are often sequentially executed in two steps. The first step involves fitting a general linear model (LM) (or a linear mixed model (LMM)) to imaging data from all subjects at each voxel and generating a statistical parametric map of test statistics (or p-values) [5,6]. The second step is to calculate adjusted p-values that account for testing the hypotheses across multiple brain regions or across many voxels of the imaging volume using various statistical methods (e.g., random field theory (RFT), false discovery rate, or permutation methods) [7,8]. Most of these methods have been implemented in existing neuroimaging software platforms, such as SPM (<http://www.fil.ion.ucl.ac.uk>), among many others.

The voxel-wise approach based on the LM (or LMM) and RFT has several obvious limitations for the analyses of imaging data, underscoring the great need for methodological development. (i) The voxel-wise approach essentially treats all voxels as independent units [9], whereas neuroimaging data are spatially correlated in nature and spatially contiguous regions of activation with rather sharp edges are usually expected. (ii) The initial smoothing step before the voxel-wise approach often blurs the image data near the edges of activated regions and thus it can dramatically increase the numbers of false positives and false negatives [11,12,13,9]. (iii) The voxel-wise approach is also based on a strong assumption that 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—an assumption that is demonstrably false.

Spatially modeling imaging data in all voxels of the 3D volume (or 2D surface) represents both computational and theoretical challenges. Spatial dependencies were commonly characterized using conditional autoregressive (CAR) or Markov random field (MRF) priors, but estimating spatial correlation for the 3D volume, in which the number of voxels ranges from ten thousands to more than 500,000 voxels, is computationally prohibited. Moreover, given the complexity of imaging data, it can be restrictive to assume parametric spatial correlation such as CAR and MRF for the whole 3D volume (or 2D surface). Another method, called ROI analysis, is to model the imaging data from all voxels within multiple regions of interest (ROIs) [14]. The ROI method based on anatomically defined ROIs only models the spatial correlation among these ROIs [14], so it essentially ignores the spatial correlation structure in the neighboring voxels within each ROI.

Moreover, the ROI method is also based on a strong assumption that all voxels in the same ROI are homogeneous, and this assumption is largely false.

This paper aims to develop and apply a multiscale adaptive regression model (MARM) for the joint analysis of neuroimaging data with behavioral and clinical variables, and then to demonstrate its superiority over the voxel-wise approach using simulated and real imaging data. The MARM is a spatial, hierarchical and adaptive procedure. The MARM builds a small sphere at each voxel and use these consecutively connected spheres across all voxels to capture local and global spatial dependence among imaging observations. The MARM also builds hierarchically nested spheres by increasing the radius of a spherical neighborhood around each voxel and combine all the data in a given radius of each voxel with appropriate weights to adaptively calculate parameter estimates and test statistics. Thus, the MARM explicitly utilizes the spatial information to carry out statistical inference, while avoiding explicitly estimating spatial correlation. The hierarchical nature of the MARM can dramatically reduce the computational complexity in computing parameter estimates. The adaptive feature of the MARM can efficiently utilize all available information in the neighboring voxels to increase the precision of parameter estimates and the power of test statistics.

The MARM represents a novel generalization of the propagation separation (PS) approach, which was originally developed for nonparametric estimation of regression curves or surfaces [11,12], in several aspects. The MARM provides a general framework for carrying out statistical inference on imaging data, whereas the PS is applied to smooth the images of parameter estimates obtained from the voxel-wise approach based on classical linear models [9]. As shown in Section 2, it is inadequate to directly use the PS approach to smooth the images of parameter estimates, which are obtained from the voxel-wise method, for most regression models, such as nonlinear regression. Compared to the parametric assumptions in the PS method for the LM, the MARM is solely based on the pseudo-likelihood function, and thus it avoids specifying any parametric distribution for imaging data. This feature is desirable for the analysis of real neuroimaging data, including brain morphological measures, because the distribution of the univariate (or multivariate) neuroimaging measurements often deviates from the Gaussian distribution [15]. We also establish the theoretical properties of the MARM, which differs substantially from those of the original PS approach, which were developed for nonparametric estimation of regression curves or surfaces based on observations from the exponential family model [12]. Particularly, we show that the MARM outperforms the voxel-wise method theoretically.

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

## 2 Multiscale Adaptive Regression Model

### 2.1 Data Structure and Model Formulation

Suppose we have 2D surfaces or 3D volumes of MRI measures and clinical variables from  $n$  subjects for  $i = 1, \dots, n$ . MRI measures might be the shape representation of the surfaces of cortical or various subcortical regions, the determinant of the Jacobian matrices based on the deformation fields estimated by the registration algorithm, functional MRI signals, or diffusion tensors and their associated invariant measures, such as fractional anisotropy [1]. Clinical variables might include pedigree information, time, demographic characteristics (e.g., age, gender, height), and diagnoses, among others. Thus, for the  $i$ -th subject, we observe an  $N_D \times 1$  vector of MRI measures, denoted by  $\mathbf{Y}_{i,\mathcal{D}} = \{Y_i(d) : d \in \mathcal{D}\}$ , and a  $k \times 1$  vector of clinical variables  $\mathbf{x}_i$ , where  $\mathcal{D}$  and  $d$ , respectively, represent a 3D volume (or 2D surface) and a voxel in  $\mathcal{D}$  and  $N_D$  equals the number of points on  $\mathcal{D}$ .

Our primary scientific interest in the analysis of neuroimaging data is to identify important brain regions to characterize the neural development of neuropsychiatric disorders, substance use disorders, and normal brains. Statistically, we often use  $\{\mathbf{Y}_{i,\mathcal{D}} : i = 1, \dots, n\}$  as responses and establish their association with a set of covariates  $\mathbf{x}_i$ , such as age and gender. This requires the specification of the conditional distribution of

$$\mathbf{Y}_{\mathcal{D}} = \{\mathbf{Y}_{i,\mathcal{D}} = (\mathbf{Y}_{i,\mathcal{D}} : i = 1, \dots, n)\}$$

given  $\mathbf{X} = \{\mathbf{x}_i : i = 1, \dots, n\}$ , that is,  $p(\mathbf{Y}_{\mathcal{D}}|\mathbf{X})$ . For MRI measures from the cross-sectional studies, it is natural to assume the independence across all subjects, that is given by

$$p(\mathbf{Y}_{\mathcal{D}}|\mathbf{X}) = \prod_{i=1}^n p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i).$$

Thus, we only need to specify  $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$  for each subject. However, even for a single observation within each cluster, the number of voxels in each brain region varies from thousands to 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 correlated in nature, and thus given the large number of voxels on each brain structure, it is statistically challenging to directly model the spatial correlations among all pairs of points [14].

The voxel-wise approach essentially assumes that

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

where  $p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d))$  is the marginal density of  $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$  or a ‘pseudo’ density function for  $Y_i(d)$  parameterized by an unknown parameter vector  $\boldsymbol{\theta}(d) =$

$(\theta_1(d), \dots, \theta_p(d))^T$  in an open subset  $\Theta$  of  $R^p$ . Note that we use the pseudo density to emphasize the possible misspecification of  $p(Y_i(d)|\mathbf{x}_i, \theta(d))$ . Model (1) comprises many statistical models such as the LM. For instance, for univariate measure, the LM assumes that

$$Y_i(d) = \mathbf{x}_i^T \beta(d) + \epsilon_i(d) \quad \text{for all } i = 1, \dots, n,$$

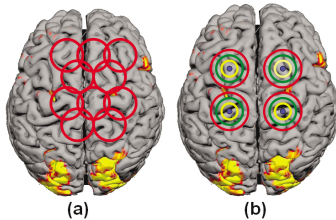
where  $\beta(d)$  is a  $(p - 1) \times 1$  regression coefficients,  $\epsilon_{i1}(d) \sim N(0, \sigma(d)^2)$ , and  $\theta(d) = (\beta(d), \sigma(d))$ . However, the linear link function  $E[Y_i(d)|\mathbf{x}_i] = \mathbf{x}_i^T \beta(d)$  and the Gaussian assumption are questionable in many applications [15]. Moreover, since the voxel-wise approach does not account for the fact that imaging data are spatially correlated and contain spatially contiguous regions of activation with rather sharp edges, it may lead to the loss of power in detecting statistical significance in the analysis of imaging data.

We formally introduce the multiscale adaptive regression model as follows. It is first assumed that for a relatively large radius  $r_0$ ,

$$p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i) \approx \prod_{d \in \mathcal{D}} p(\{Y_i(d') : d' \in N(d, r_0)\}|\mathbf{x}_i), \tag{2}$$

where  $N(d, r_0)$  denotes the set of all voxels in a spherical neighborhood of a voxel  $d$  with radius  $r_0$ . That is, we can approximate the joint distribution of  $\mathbf{Y}_{i,\mathcal{D}}$  by the product of the joint distributions of  $\{Y_i(d') : d' \in N(d, r_0)\}$ . Using data in  $N(d, r_0)$  for relatively large  $r_0$  preserves the neighboring correlation structure in the imaging data (see the panel (a) in Figure 1 for an illustration). Moreover, since the spherical neighborhoods for all voxels are consecutively connected, equation (2) can capture a substantial amount of spatial information in the imaging data. Note that the right hand-side of equation (2) is essentially a composite likelihood [16,17].

Second, we consider the specification of  $p(\{Y_i(d') : d' \in N(d, r_0)\}|\mathbf{x}_i)$ . Since our primary interest is to make statistical inference about  $\theta(d)$ , we avoid specifying spatial correlations among all the  $\{Y_i(d') : d' \in N(d, r_0)\}$ . Instead, we assume that  $p(\{Y_i(d') : d' \in N(d, r_0)\}|\mathbf{x}_i)$  can be approximated by



**Fig. 1.** Illustrating the key features of the multiscale adaptive regression model. For a relatively large radius  $r_0$ , panel (a) shows the spherical neighborhoods  $N(d, r_0)$  of multiple points  $d$  on the cortical surface. Panel (b) shows the spherical neighborhoods with four different bandwidths  $h$  of the four selected points  $d$  on the cortical surface.

$$p(\{Y_i(d') : d' \in N(d, r_0)\} | \mathbf{x}_i) \approx \left\{ \prod_{d' \in N(d, r_0)} p(Y_i(d') | \mathbf{x}_i, \boldsymbol{\theta}(d'))^{\omega(d, d'; r_0)} \right\}, \quad (3)$$

where  $\omega(d, d'; h)$  is a weight function of two voxels and a radius  $h$  that characterizes the similarity between the data in voxels  $d$  and  $d'$ . We require that  $\omega(d, d'; h)$  be independent of  $i$  just for simplicity. In imaging data, voxels, which are not on the boundary of regions of activation, often have a neighborhood in which  $\boldsymbol{\theta}(d)$  is nearly constant. This assumption reflects the fact that imaging data are spatially correlated and contain spatially contiguous regions of activation with rather sharp edges. Incorporating this assumption leads to

$$p(\{Y_i(d') : d' \in N(d, r_0)\} | \mathbf{x}_i) \approx \left\{ \prod_{d' \in N(d, r_0)} p(Y_i(d') | \mathbf{x}_i, \boldsymbol{\theta}(d))^{\omega(d, d'; r_0)} \right\}. \quad (4)$$

Equation (4) allows us to combine all data in  $N(d, r_0)$  to make inference about  $\boldsymbol{\theta}(d)$ , which can substantially increase the efficiency in estimating  $\boldsymbol{\theta}(d)$ . Moreover, the weights  $\omega(d, d'; r_0)$  can prevent incorporating voxels whose data do not contain information on  $\boldsymbol{\theta}(d)$ , and thus preserve the edges of the regions of activation.

An important question that we need to address is how to determine  $\omega(d, d'; r_0)$ . We use a multiscale strategy to adaptively determine  $\{\omega(d, d'; r_0) : d, d' \in \mathcal{D}\}$  and estimate  $\boldsymbol{\theta}(d)$ . Specifically, we select a sequence of bandwidths  $h_0 = 0 < h_1 < \dots < h_S = r_0$  ranging from the smallest scale  $h_0 = 0$  to the largest scale  $h_S = r_0$ . By setting  $\omega(d, d'; h_0 = 0) = 1$ , we can estimate  $\boldsymbol{\theta}(d)$  at scale  $h_0$ , denoted by  $\tilde{\boldsymbol{\theta}}(d; h_0 = 0)$ , and then we use some methods as detailed below to calculate  $\omega(d, d'; h_1)$  at scale  $h_1$  based on  $\{\tilde{\boldsymbol{\theta}}(d; h_0 = 0) : d \in \mathcal{D}\}$ . In this way, we can sequentially determine  $\omega(d, d'; h_s)$  and adaptively update  $\tilde{\boldsymbol{\theta}}(d; h_s)$  from  $h_0 = 0$  to  $h_S = r_0$  (see the panel (b) of Figure 1 for an illustration). A path diagram is given below:

$$\begin{array}{ccccccc}
 \omega(d, d'; h_0) & \omega(d, d'; h_1) & \dots & \omega(d, d'; h_S = r_0) & & & \\
 \downarrow & \nearrow & \downarrow & \nearrow & \dots & \nearrow & \downarrow \\
 \tilde{\boldsymbol{\theta}}(d; h_0) & \tilde{\boldsymbol{\theta}}(d; h_1) & \dots & \tilde{\boldsymbol{\theta}}(d; h_S) & & & 
 \end{array} \quad (5)$$

At each iteration, the computation involved for the MARM is of the same order as that for the voxel-wise approach. Thus, this multiscale method provides an efficient method for adaptively exploring the neighboring areas of each voxel. Since this multiscale method sequentially includes more data at each voxel, it will adaptively increase the statistical efficiency in estimating  $\boldsymbol{\theta}(d)$  in a homogenous region and decreases the variation of the weights  $\omega(d, d'; h)$ . This multiscale method distinguishes MARM from the composite likelihood methods proposed in the literature [16,17].

### 2.2 Estimation and Hypothesis Testing at a Fixed Scale

At a fixed scale  $h$ , we consider the weighted maximum likelihood estimates of  $\boldsymbol{\theta}(d)$  across all voxels  $d \in \mathcal{D}$  for given weights  $\omega(d, d'; h)$ . The weighted quasi-likelihood function  $\ell_n(\boldsymbol{\theta}(d); h, \omega)$  is given by

$$\ell_n(\boldsymbol{\theta}(d); h, \omega) = \sum_{i=1}^n \sum_{d' \in N(d, h)} \omega(d, d'; h) \log p(Y_i(d') | \mathbf{x}_i, \boldsymbol{\theta}(d)). \quad (6)$$

The maximum weighted quasi-likelihood (MWQL) estimate of  $\boldsymbol{\theta}$  is

$$\hat{\boldsymbol{\theta}}(d, h) = \operatorname{argmax}_{\boldsymbol{\theta}(d)} n^{-1} \ell_n(\boldsymbol{\theta}(d); h, \omega). \quad (7)$$

We use the Newton-Raphson algorithm to calculate  $\hat{\boldsymbol{\theta}}(d, h)$  by iterating

$$\hat{\boldsymbol{\theta}}(d, h)^{(t+1)} = \hat{\boldsymbol{\theta}}(d, h)^{(t)} + \{-\partial_{\hat{\boldsymbol{\theta}}(d)}^2 \ell_n(\hat{\boldsymbol{\theta}}(d, h)^{(t)}; h, \omega)\}^{-1} \partial_{\hat{\boldsymbol{\theta}}(d)} \ell_n(\hat{\boldsymbol{\theta}}(d, h)^{(t)}; h, \omega),$$

where  $\partial_{\boldsymbol{\theta}(d)}$  and  $\partial_{\hat{\boldsymbol{\theta}}(d)}^2$  denote, respectively, the first- and second-order partial derivatives with respect to  $\boldsymbol{\theta}(d)$  evaluated at  $\hat{\boldsymbol{\theta}}(d, h)^{(t)}$ . In practice, to stabilize the Newton-Raphson algorithm, we may approximate  $-\partial_{\hat{\boldsymbol{\theta}}(d)}^2 \ell_n(\hat{\boldsymbol{\theta}}(d, h)^{(t)}; h, \omega)$  by  $E[-\partial_{\hat{\boldsymbol{\theta}}(d)}^2 \ell_n(\hat{\boldsymbol{\theta}}(d, h)^{(t)}; h, \omega)]$ . The Newton-Raphson algorithm stops when the absolute difference between consecutive  $\hat{\boldsymbol{\theta}}(d, h)^{(t)}$ s is smaller than a predefined small number, say  $10^{-4}$ . After convergence,  $\operatorname{Cov}[\hat{\boldsymbol{\theta}}(d, h)] = \Sigma_n(\hat{\boldsymbol{\theta}}(d, h))$  can be approximated by  $[\Sigma_{n,1}(\hat{\boldsymbol{\theta}}(d, h))]^{-1} \Sigma_{n,2}(\hat{\boldsymbol{\theta}}(d, h)) [\Sigma_{n,1}(\hat{\boldsymbol{\theta}}(d, h))]^{-1}$ , where  $\Sigma_{n,1}(\boldsymbol{\theta}) = -\partial_{\boldsymbol{\theta}(d)}^2 \ell_n(\boldsymbol{\theta}; h, \omega)$  and  $\Sigma_{n,2}(\boldsymbol{\theta}) = \sum_{i=1}^n [\sum_{d' \in N(d, h)} \omega(d, d'; h) \partial_{\boldsymbol{\theta}(d)} \log p(Y_i(d') | \mathbf{x}_i, \boldsymbol{\theta})]^{\otimes 2}$ , in which  $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$  for any vector  $\mathbf{a}$ .

Our choice of which hypotheses to test was motivated by either a comparison of brain structure across diagnostic groups or the detection of a change in brain structure across time [1]. These questions usually can be formulated as the testing of linear hypotheses about  $\boldsymbol{\theta}(d)$

$$H_{0,\mu} : R\boldsymbol{\theta}(d) = \mathbf{b}_0 \quad \text{vs.} \quad H_{1,\mu} : R\boldsymbol{\theta}(d) \neq \mathbf{b}_0, \quad (8)$$

where  $\mu = R\boldsymbol{\theta}(d)$ ,  $R$  is a  $r \times k$  matrix of full row rank and  $\mathbf{b}_0$  is an  $r \times 1$  specified vector. We test the null hypothesis  $H_{0,\mu} : R\boldsymbol{\theta}(d) = \mathbf{b}_0$  using the score test statistic

$$W_\mu(d, h) = [R\hat{\boldsymbol{\theta}}(d, h) - \mathbf{b}_0]^T [R\hat{\Sigma}_n(\hat{\boldsymbol{\theta}}(d; h))R^T]^{-1} [R\hat{\boldsymbol{\theta}}(d, h) - \mathbf{b}_0]. \quad (9)$$

To test whether  $H_{0,\mu}$  holds in all voxels of the region under study, we consider the false discovery rate (FDR) method [10].

### 2.3 Adaptive Estimation and Testing Procedure

We develop an adaptive estimation and testing (AET) procedure for MARM. The AET procedure starts with a single voxel  $d$  and then successively increases the radius (or bandwidth)  $h$  of a spherical neighborhood around  $d$ . Each voxel  $d'$  in the neighborhood of  $d$  will be given a weight  $\omega(d, d'; h_s)$  that depends on the distance between  $d$  and  $d'$  and the similarity between  $\hat{\boldsymbol{\theta}}(d, h_{s-1})$  and  $\hat{\boldsymbol{\theta}}(d', h_{s-1})$ . Then, we use all the data in a given neighborhood of  $d$  with bandwidth  $h_s$  and the weight in each of these voxels to obtain updated estimates  $\hat{\boldsymbol{\theta}}(d, h_s)$  and

$W_\mu(d, h_s)$  at  $d$ , respectively. Finally, we use a sequence of  $\hat{\boldsymbol{\theta}}(d, h_s)$  and  $W_\mu(d, h_s)$  as a function of  $h$  to construct the final estimate for  $\boldsymbol{\theta}(d)$  and calculate the final test statistic  $W_\mu(d)$  for testing hypotheses on  $\boldsymbol{\theta}(d)$  at  $d$ .

The AET procedure consists of five key steps as follows.

In the initialization step (i), we generate a geometric series  $\{h_k = c_h^s : s = 1, \dots, S\}$  of bandwidths with  $h_0 = 0$ , where  $c_h$  is a number in  $(1, 2)$ , say  $c_h = 1.25$ . At each voxel  $d$ , we calculate  $\hat{\boldsymbol{\theta}}(d, h_0)$  and  $W_\mu(d, h_0)$ , which are the same as those from the voxel-wise approach. We then set  $s = 1$ , and  $h_1 = c_h$ .

In the weights adaptation step (ii), we compute adaptive weights

$$\omega(d, d'; h_s) = K_{loc}(\|d - d'\|_2/h_s)K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_{s-1})), \tag{10}$$

where  $K_{loc}(\cdot)$  and  $K_{st}(\cdot)$  are two kernel functions with compact support,  $\|\cdot\|_2$  denotes the Euclidean norm, and  $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$  denotes a weighted function based on the estimates of  $\{\boldsymbol{\theta}(d) : d \in \mathcal{D}\}$  at the  $(s - 1)$ th iteration. The adaptive weights can downweight voxels  $d'$  in  $\ell_n(\boldsymbol{\theta}(d); h, d)$  if  $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$  is large.

In the estimation step (iii), we calculate  $\hat{\boldsymbol{\theta}}(d, h_s)$  and  $W_\mu(d, h_s)$ , which are defined in equations (7) and (10), respectively, at voxel  $d$  for the  $s$ th scale.

In the memory step (iv), we set  $\tilde{\boldsymbol{\theta}}(d, h_0) = \hat{\boldsymbol{\theta}}(d, h_0)$  and then update  $\tilde{\boldsymbol{\theta}}(d, h_s)$  for  $s > 0$  as

$$\tilde{\boldsymbol{\theta}}(d, h_s) = \hat{\boldsymbol{\theta}}(d, h_s)\eta_s/s + (1 - \eta_s/s)\tilde{\boldsymbol{\theta}}(d, h_{s-1}), \tag{11}$$

where  $\eta_s = K_{loc}(D_{\boldsymbol{\theta}}(d, h_s)/C_0) \in (0, 1)$  and

$$D_{\boldsymbol{\theta}}(d, h_s) = [\hat{\boldsymbol{\theta}}(d, h_s) - \hat{\boldsymbol{\theta}}(d, h_0)]^T \text{Cov}(\hat{\boldsymbol{\theta}}(d, h_0))^{-1} [\hat{\boldsymbol{\theta}}(d, h_s) - \hat{\boldsymbol{\theta}}(d, h_0)]. \tag{12}$$

The  $D_{\boldsymbol{\theta}}(d, h_s)$  measures the difference between  $\hat{\boldsymbol{\theta}}(d, h_s)$  and  $\hat{\boldsymbol{\theta}}(d, h_0)$  at the same voxel  $d$ . Finally, we calculate an estimator of the covariance matrix of  $\tilde{\boldsymbol{\theta}}(d, h_s)$ , denoted by  $\text{Cov}(\tilde{\boldsymbol{\theta}}(d, h_s))$  and compute the Wald test statistic as

$$\tilde{W}_\mu(d, h) = [R\tilde{\boldsymbol{\theta}}(d, h_s) - \mathbf{b}_0]^T \text{Cov}(\tilde{\boldsymbol{\theta}}(d, h_s))^{-1} [R\tilde{\boldsymbol{\theta}}(d, h_s) - \mathbf{b}_0]. \tag{13}$$

In the stopping step (v), when  $s = S$ , we compute the  $p$ -values for  $\tilde{W}_\mu(d, h)$ , apply FDR to detect significant voxels and then stop, otherwise set  $h_{s+1} = c_h h_s$ , increase  $s$  by 1 and continue with the weight adaptation step (ii). The maximal step  $S$  can be taken to be relatively small, say 6, such that the largest spherical neighborhood of each voxel only contains a relatively small number of voxels compared with the whole volume.

*Remark 1.* The memory step in equation (11) differs substantially from that in the PS approach [12]. Equation (11) is exactly a stochastic approximation algorithm [18]. The sequence  $\{1/s : s = 1, \dots\}$  is introduced to cancel out the noise introduced in each iteration. Putting more weight  $1/s$  at the beginning is very appealing in imaging analysis, because use of a local approximation often decreases the estimation error in the first few steps of the procedure and starts to slowly increase the estimation error as  $h_s$  gets large. In addition, compared with the memory step in the PS approach, we use the distance between  $\hat{\boldsymbol{\theta}}(d, h_s)$

and  $\hat{\boldsymbol{\theta}}(d, h_0)$  to control the estimation error of  $\tilde{\boldsymbol{\theta}}(d, h_0)$ . Since  $\hat{\boldsymbol{\theta}}(d, h_0)$  is a  $\sqrt{n}$  consistent estimate of  $\boldsymbol{\theta}(d)$ ,  $\eta_s = K_{loc}(D_{\boldsymbol{\theta}}(d, h_s)/C_0)$  ensures that  $\boldsymbol{\theta}(d, h_s)$  is also a  $\sqrt{n}$  consistent estimate of  $\boldsymbol{\theta}(d)$  for all  $s > 0$ .

*Remark 2.* There is an efficient way for selecting the initial value  $\boldsymbol{\theta}(d, h_s)^{(0)}$  for the Newton-Raphson algorithm by setting  $\boldsymbol{\theta}(d, h_s)^{(0)} = \hat{\boldsymbol{\theta}}(d, h_{s-1})$  for each  $s > 0$ . Since the AET procedure always downweights voxels  $d'$  in  $\ell_n(\boldsymbol{\theta}(d); h, d)$  if  $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$  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  $\boldsymbol{\theta}(d, h_s)^{(0)} = \hat{\boldsymbol{\theta}}(d, h_{s-1})$ , the Newton-Raphson algorithm converges very fast, and thus the additional computation involved for the MARM is very light compared to the voxel-wise approach.

### 3 Simulation Study

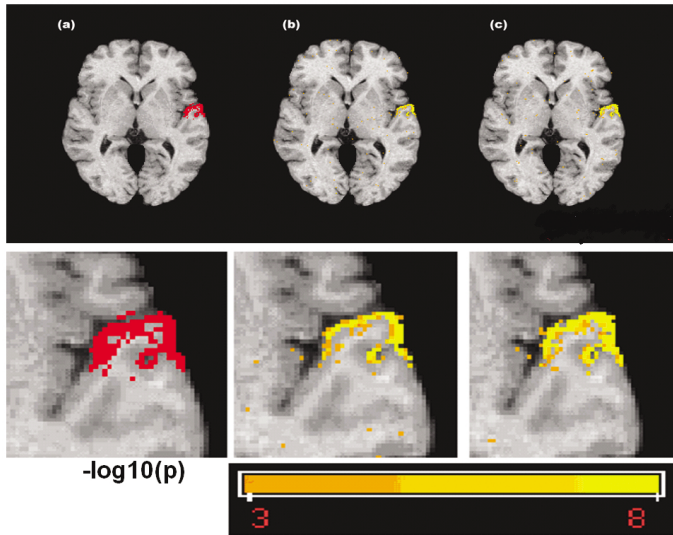
We only presented one set of Monte Carlo simulations to examine the finite sample performance of  $\tilde{W}_{\mu}(d, h)$  with respect to different scales  $h$  at the levels of a single voxel and an entire brain region with known ground truth.

We applied a simulation model to automatically simulate realistic intra-individual deformations associated with tissue atrophy or growth on brain images from two groups [19]. We chose a specified location and a fixed radius in the white matter and then simulated spherical atrophy for all 20 subjects in each group (see Figure 2). The growth rates for each subject in the first and second groups were generated from  $N(0.95, 0.01)$  and  $N(1, 0.01)$ , respectively.

We used simulated deformations and images with the known ground truth to demonstrate the superiority of the MARM over the voxel-wise approach. The true deformation area was highlighted in red (see the panel (a) of Figure 2). We applied the MARM with  $c_h = 1.25$ ,  $S = 6$  and computed the  $p$ -values of  $\tilde{W}_{\mu}(d, h)$  across the 3D volume at each iteration. Note that the results obtained from  $h_0 = 0$  correspond to those from the voxel-wise approach. Our results show a clear advantage of the MARM in detecting an accurate group difference as we increase the bandwidth  $h$  of the spherical neighborhood (compare the panels (b) and (c) of Figure 2). We calculated the ratio of voxels within the true deformation regions, whose  $p$ -values are smaller than 0.0001. The ratios for  $h_0$  and  $h_6$  are 49% and 68%, respectively. That is, the MARM based on  $h_6$  leads to 19% improvement compared with the traditional voxel-wise approach with  $h_0 = 0$ .

### 4 Real Data Analysis

Alzheimer's disease (AD) is the most common form of dementia in people over 65 years of age. MRI has been used to develop imaging-based biomarkers for AD, measure spatial patterns of atrophy, and their evolution with disease progressions. We used a subset of a large MRI dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([www.loni.ucla.edu/ADNI](http://www.loni.ucla.edu/ADNI)). Our dataset includes 90 subjects, including 45 cognitively normal individuals

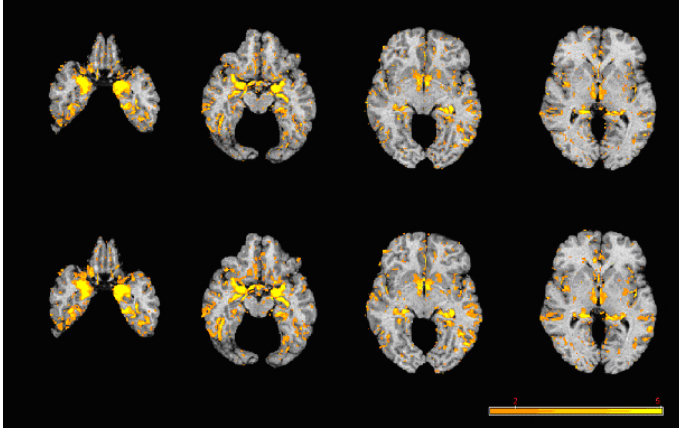


**Fig. 2.** Voxel-wise analysis of group difference. From left to right in the first row, it shows the true deformation region in red in panel (a), the raw  $-\log_{10}(p)$  values of the Wald test statistics  $\tilde{W}_\mu(d, h_0)$  in panel (b), and the raw  $-\log_{10}(p)$  values of the Wald test statistics  $\tilde{W}_\mu(d, h_6)$  based on a  $\chi^2$  distribution in panel (c). The second row shows the enlarged deformation regions of the corresponding figures in the first row.

(CN) (mean age S.D., 77.07 3.89), and 45 AD patients (77.32 6.01). The mini mental state examination (MMSE) scores (mean S.D.) of each group at baseline were 29.16 0.92, and 23.13 1.75, respectively. The two groups were relatively well-balanced in terms of gender (23,25 women in each of the 2 groups, respectively). The imaging data include standard T1-weighted MR images acquired sagittally using volumetric 3D MPRAGE with  $1.25 \times 1.25 \text{ mm}^2$  in-plane spatial resolution and 1.2 mm thick sagittal slices (8 flip angle).

The T1-weighted MRIs were preprocessed in six consecutive steps. These steps included (i) alignment to the AC-PC plane; (ii) removal of extra-cranial material (skull-stripping); (iii) tissue segmentation into grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using a brain tissue segmentation method proposed; (iv) high-dimensional image warping to a standardized coordinate system, a brain atlas (template) that was aligned with the MNI coordinate space; (vi) formation of regional volumetric maps, named RAVENS maps, for GM, WM, and CSF using tissue preserving image warping.

We identify the spatial patterns of brain atrophy in Alzheimer's disease (AD) via the analysis of the RAVENS maps of GM and WM obtained from the ANDI dataset. To control for the effects of covariates (diagnosis, age, weight, and gender), we considered model  $y_i(d) = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i(d)$  for respective RAVENS maps at each voxel. The  $\mathbf{x}_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i})^T$  is a  $4 \times 1$  vector, in which  $x_{1i}$  is Age/10,  $x_{2i}$  is gender,  $x_{4i}$  denotes the weight, and  $x_{4i}$  denotes the diagnosis (1



**Fig. 3.** Voxel-based analysis of group difference between CN and AD based on the raw  $-\log_{10}(P)$  values of the Wald test statistics. Four selected slices are presented. The first and second rows represent the results from the multiscale LM with  $h_0 =$  and  $h_6$ , respectively.

AD and 0 CN). We applied the AET procedure with  $c_h = 1.25$  and  $S = 6$  to carry out the statistical analysis. Figure 2 shows a clear advantage of the MARM in detecting more significant and smoothly area for the group differences between CN and AD as the bandwidth  $h$  increases. We observed the significant difference between CN and AD in the hippocampus and the entorhinal cortex.

## 5 Discussion

We have developed the MARM for spatial and adaptive analysis of imaging data. We have used simulation studies and real data to show that the MARM significantly outperforms the classical voxel-wise approach. Many issues still merit further research.

**Acknowledgments.** We thank the reviewers for their thoughtful comments. This work was supported in part by NSF grants SES-06-43663 and BCS-08-26844 and NIH grants UL1-RR025747-01 and R21AG033387 to Dr. Zhu, NIH grants GM 70335 and CA 74015 to Dr. Ibrahim, NIH grant R01NS055754 to Dr. Lin, and NIH grant R01EB006733 and R03EB008760 to Dr. Shen.

## References

1. Thompson, P.M., Cannon, T.D., Toga, A.W.: Mapping Genetic Influences on Human Brain Structure. *The Annals of Medicine* 24, 523–536 (2002)
2. Friston, K.J.: *Statistical Parametric Mapping: the Analysis of Functional Brain Images*. Academic Press, London (2007)

3. Rogers, B.P., Morgan, V.L., Newton, A.T., Gore, J.C.: Assessing Functional Connectivity in the Human Brain by fMRI. *Magnetic Resonance Imaging* 25(10), 1347–1357 (2007)
4. Huettel, S.A., Song, A.W., McCarthy, G.: *Functional Magnetic Resonance Imaging*. Sinauer Associates, Inc. (2004)
5. Friston, K., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak, R.S.J.: Statistical Parametric Maps in Functional Imaging: a General Linear Approach. *Human Brain Mapping* 2, 189–210 (1995)
6. Beckmann, C.F., Jenkinson, M., Smith, S.M.: General Multilevel Linear Modeling for Group Analysis in fMRI. *NeuroImage* 20, 1052–1063 (2003)
7. Nichols, T., Hayasaka, S.: Controlling the Family-wise Error Rate in Functional Neuroimaging: a Comparative Review. *Statistical Methods in Medical Research* 12, 419–446 (2003)
8. Worsley, K.J., Taylor, J.E., Tomaiuolo, F., Lerch, J.: Unified Univariate and Multivariate Random Field Theory. *NeuroImage* 23, 189–195 (2004)
9. Tabelow, K., Polzehl, J., Voss, H.U., Spokoiny, V.: Analyzing fMRI Experiments with Structural Adaptive Smoothing Procedures. *NeuroImage* 33, 55–62 (2006)
10. 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* 57, 289–300 (1995)
11. Polzehl, J., Spokoiny, V.G.: Image Denoising: Pointwise Adaptive Approach. *Annals of Statistics* 31, 30–57 (2003)
12. Polzehl, J., Spokoiny, V.G.: Propagation-Separation Approach for Local Likelihood Estimation. *Probab. Theory Relat. Fields* 135, 335–362 (2006)
13. Qiu, P.: *Image Processing and Jump Regression Analysis*. John Wiley & Sons, New York (2005)
14. Bowman, F.D.: Spatio-temporal Models for Region of Interest Analyses of Functional Mapping Experiments. *Journal of American Statistical Association* 102, 442–453 (2007)
15. Luo, W., Nichols, T.: Diagnosis and Exploration of Massively Univariate fMRI Models. *NeuroImage* 19, 1014–1032 (2003)
16. Lindsay, B.G.: Composite Likelihood Methods. *Contemp. Math.* 80, 221–240 (1988)
17. Varin, C.: On Composite Marginal Likelihoods. *Advances in Statistical Analysis* 92, 1–28 (2008)
18. Robbins, H., Monro, S.: A Stochastic Approximation Method. *Annals of Mathematical Statistics* 22, 400–407 (1951)
19. Xue, Z., Shen, D., Karacali, B., Stern, J., Rottenberg, D., Davatzikos, C.: Simulating Deformations of MR Brain Images for Validation of Atlas-based Segmentation and Registration Algorithms. *NeuroImage* 33, 855–866 (2006)