Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2011)
73, *Part* 4, *pp.* 559–578

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, USA

[Received May 2009. Final revision November 2010]

Summary. Neuroimaging studies aim to analyse imaging data with complex spatial patterns in a large number of locations (called voxels) on a two-dimensional surface or in a three-dimensional 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 extent of smoothing and low statistical power in detecting spatial patterns. We propose a multiscale adaptive regression model to integrate the propagation–separation approach with statistical modelling at each voxel for spatial and adaptive analysis of neuroimaging data from multiple subjects. The multiscale adaptive regression model has three features: being spatial, being hierarchical and being adaptive. We use a multiscale adaptive estimation and testing procedure to utilize imaging observations from the neighbouring voxels of the current voxel to calculate parameter estimates and test statistics adaptively. 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 the multiscale adaptive regression model 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 have been or are being widely conducted to collect neuroimaging data including anatomical and functional images from multiple subjects to understand better 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. the hippocampus) are extracted for understanding 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 fibre tracts are extracted for quantitatively assessing the integrity of anatomical white matter 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: Hongtu Zhu, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420, USA.
E-mail: hzhu@bios.unc.edu

 1369–7412/11/73559

Following spatial normalization, imaging observations for each subject are observed in a large number of locations (called voxels), that number in the thousands to millions, on a common two-dimensional surface or in a common three-dimensional volume. Conventional analyses of high dimensional imaging data are often executed in two sequential steps: spatially smoothing the imaging data and then independently fitting a statistical model, such as a general linear model, at each voxel, which is called a voxelwise method. Most smoothing methods are independent of the imaging data and apply the same amount of smoothness throughout the whole image. See, for example, Yue *et al.* (2010) for overviews of smoothing methods that are 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 smooth images from a single subject adaptively and spatially. Tabelow *et al.* (2006, 2008a, b, c) used the original PS idea to develop a multiscale adaptive linear model to denoise FMRI and diffusion tensor images from a single subject adaptively and spatially.

The existing voxelwise methods for analysing high dimensional data involve fitting a statistical model, such as a linear model, 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). These voxelwise 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), voxelwise methods can suffer from the arbitrary choice of smoothing extent in the initial smoothing step and thus dramatically increase the number of false positive and false negative results. Furthermore, as pointed out by Worsley (2003) and Tabelow *et al.* (2006), voxelwise 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 in many neuroimaging studies.

Spatially modelling neuroimaging data in the three-dimensional volume (or two-dimensional surface) represents both computational and theoretical challenges. It is common to use conditional auto-regressive, Markov random-field and other spatial correlation priors to characterize spatial dependence between spatially connected voxels (Besag, 1986; Banerjee *et al.*, 2004). However, calculating the normalizing factor of Markov random fields and estimating spatial correlation for a large number of voxels in the three-dimensional volume (or two-dimensional surface) are computationally prohibitive (Zhu *et al.*, 2007a; Bowman, 2007). Moreover, it can be restrictive to assume a specific type of correlation structure, such as conditional auto-regressive and Markov random field, for the whole three-dimensional volume (or two-dimensional surface).

The goal of this paper is to develop a multiscale adaptive regression model (MARM) for the spatial and adaptive analysis of neuroimaging data. The MARM integrates the PS approach and voxelwise methods and thus it is a generalization of the PS approach (Polzehl and Spokoiny, 2000, 2006) to neuroimaging data from multiple subjects. The MARM has three features: being spatial, being hierarchical and being adaptive. It 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. Owing to its hierarchical and adaptive nature, the 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.

The 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 the MARM as the number of subjects (or images) increases to ∞. The covariance estimate of the adaptive estimator in the MARM has a simple form. Our new theoretical results show that, in the 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 methodology proposed, 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 fitted a multivariate linear model with age and age^2 as covariates and calculated the Wald statistics and their associated p-values for testing an age-dependent effect. Inspecting Figs 1(a)–1(c) reveals that the size of significant regions

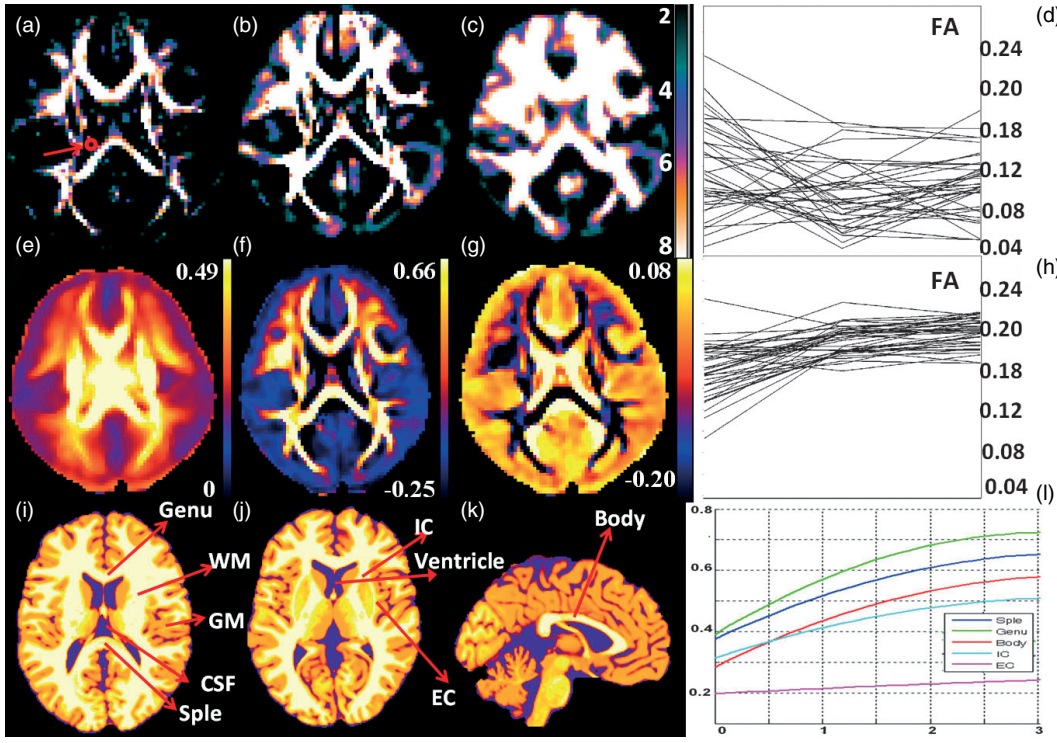


Fig. 1. Results from the neonatal project on brain development: (a) Bonferroni-corrected $-\log_{10}(p)$ values of $W_\mu(d, h_0)$ from a selected slice and a selected voxel in the red circle in the ventricle; (b) Bonferroni-corrected $-\log_{10}(p)$ values of $W_\mu(d, h_{10})$ from the same selected slice; (c) Bonferroni-corrected $-\log_{10}(p)$ values of the Wald test statistics obtained from the Gaussian-kernel-smoothed FA images for the same selected slice; (d) longitudinal trajectories of unsmoothed FA values in the red voxel identified in (a); (e) estimated $\hat{\beta}_1(d, h_{10})$; (f) estimated $\hat{\beta}_2(d, h_{10})$; (g) estimated $\hat{\beta}_3(d, h_{10})$; (h) longitudinal trajectories of the Gaussian-kernal-smoothed FA values in the red voxel in panel (a); (i)–(k) anatomical images with eight labelled 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); (l) growth patterns from the regions of interest located in the *splenium*, *genu* and body of *corpus callosum*, internal capsule and external capsule for FA

and degree of significance that are 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 analysed the same FA data set using the MARM and tested the age-dependent effect across all voxels. The MARM can preserve the edges of significant regions compared with the results from the smoothed images (Figs 1(b) and 1(c)). In contrast, the significant regions based on the smoothed images even spread over cerebrospinal fluid areas (Fig. 1(c)), in which FA values should be close to 0 and have no age-dependent effect. In Section 4, we shall revisit this data set.

Section 2 of this paper presents the 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 the 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 the MARM. Section 4 illustrates an application of the proposed methods in a real neuroimaging data set. We present concluding remarks in Section 5.

The programs that were used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Multiscale adaptive regression model

2.1. Model formulation

We consider imaging measurements in the three-dimensional volume (or on the two-dimensional surface) and clinical variables from n subjects. Without loss of generality, we focus on the three-dimensional volume. Let $\mathcal{D}$ and d respectively represent a three-dimensional 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 ith 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 $\mathcal{D}$, which is 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 $\mathbf{Y}_{\mathcal{D}} = \{\mathbf{Y}_{i,\mathcal{D}} : i = 1, \ldots, n\}$ given $\mathbf{X} = \{\mathbf{x}_i : i = 1, \ldots, n\}$, i.e. $p(\mathbf{Y}_{\mathcal{D}}|\mathbf{X})$. For a cross-sectional design, it is natural to assume that data from different subjects are independent, i.e.

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

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 totalling 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 between all pairs of points simultaneously.

The voxelwise approach essentially assumes that

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

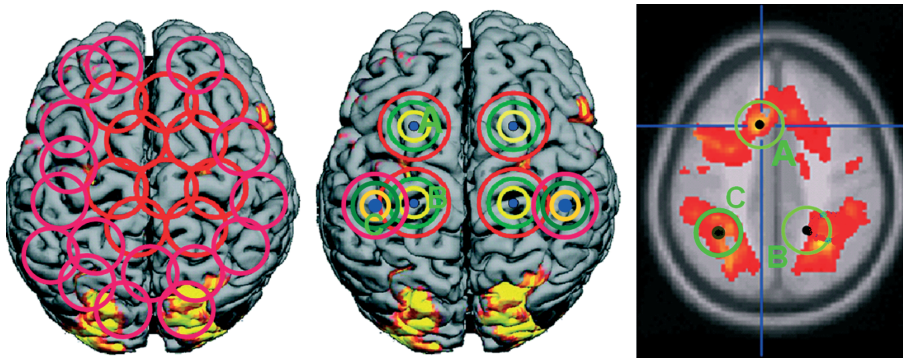
where $p\{Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d)\}$ is the marginal density of $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$ at voxel d and $\boldsymbol{\theta}(d) = (\theta_1(d), \ldots,$

$\theta_p(d))^{\mathrm{T}}$ is a $p \times 1$ vector in an open subset Θ of R^p, in which p is an integer. Moreover, the voxelwise approach makes a strong perfect registration assumption, i.e., 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. Owing 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 sufficiently general to comprise most statistical models including linear models in the neuroimaging literature. However, since the voxelwise 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 MARM 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 neighbouring voxels of d to estimate $\boldsymbol{\theta}(d)$. Similarly to standard kernel smoothing methods (Qiu, 2005), we consider a spherical neighbourhood of d with a radius (or bandwidth) r_0, which is denoted by $B(d, r_0)$. By assuming spatial independence between $\{Y_i(d') : d' \in B(d, r_0)\}$, we construct a weighted likelihood to estimate $\boldsymbol{\theta}(d)$, which is denoted by $p_{\mathrm{W}}\{Y_i(d') : d' \in B(d, r_0)|\mathbf{x}_i, \boldsymbol{\theta}(d)\}$, as follows:

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

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. 2(c)) often have a neighbourhood in which $\boldsymbol{\theta}(d)$ is nearly constant. In this case, $\omega(d, d'; h)$ for voxel d' in the neighbourhood of voxel d is greater than 0 and thus $p_{\mathrm{W}}\{\boldsymbol{\theta}(d)|Y_i(d') : d' \in B(d, r_0)\}$ allows borrowing 'good' information from these neighbouring voxels. Furthermore, we assume that $\omega(d, d'; h)$ is independent of i just for notational simplicity.



(a) (b) (c)

Fig. 2. Illustration of the key features in the MARM: for a relatively large radius r_0 (a) shows the overlapping spherical neighbourhoods $B(d, r_0)$ of multiple points (or voxels) d on the cortical surface, (b) shows the spherical neighbourhoods with four different bandwidths h of the six selected points d on the cortical surface and (c) shows the spherical neighbourhoods $B(d, r_0)$ of three selected voxels in a three-dimensional volume, in which voxels A and C are inside the activated regions, whereas voxel B is on the boundary of an activated region

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 spatial independence between imaging data, we take the product of $p_{\mathrm{W}}\{Y_i(d') : d' \in B(d, r_0) | \mathbf{x}_i, \boldsymbol{\theta}(d)\}$ for all $d \in \mathcal{D}$ and then obtain a weighted likelihood function of the MARM for $\mathbf{Y}_{i,\mathcal{D}}$ given by

$$p_{\mathrm{W}}(\mathbf{Y}_{i,\mathcal{D}} | \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\omega}) = \prod_{d \in \mathcal{D}} \left[\prod_{d' \in B(d, r_0)} p\{Y_i(d') | \mathbf{x}_i, \boldsymbol{\theta}(d)\}^{\omega(d, d'; r_0)} \right]. \tag{3}$$

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

2.2. Examples

The 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, the MARM reduces to the PS approach. For illustration, we consider the following three examples.

2.2.1. Example 1

We consider a multivariate non-linear model at each voxel given by

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

for $i = 1, \ldots, n$ and $d \in \mathcal{D}$, where $\mu(\cdot, \cdot)$ is a known $m \times 1$ vector of non-linear 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 0 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 spatial independence between imaging data, then $\log\{p_{\mathrm{W}}(\mathbf{Y}_{i,\mathcal{D}} | \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\omega})\}$ based on model (4) is given by

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

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 multiscale adaptive multivariate linear model for analysis of neuroimaging data (Tabelow *et al.*, 2006, 2008a, b, c).

2.2.2. Example 2

We consider a generalized linear model 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

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

where $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions. Moreover, $\eta_i\{\boldsymbol{\beta}(d)\} = \eta[g\{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}(d)\}]$ 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 the MARM under spatial independence is given by

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

2.2.3. Example 3

In an FMRI session, n FMRI volumes are acquired at acquisition times $t_1, \ldots, t_n$ while a subject performs a cognitive or behavioural 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 0 and variance $1/\tau(d)$ and $\mathbf{x}_i$ may include responses to differing types of stimulus, 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, artefacts from the differing magnetic field susceptibilities of neighbouring tissues, non-rigid motion and preprocessing methods (registration or normalization) among many others (Huettel *et al.*, 2004; Lazar, 2008). By performing a prewhitening procedure, we may assume that $\{\varepsilon_i(d) : i = 1, \ldots, n\}$ 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 $\boldsymbol{\theta}(d) = (\beta(d), \tau(d))$, then the weighted quasi-likelihood function of the 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. The MAET procedure 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. The MAET procedure starts with building a sequence of nested spheres with increasing radii $h_0 = 0 < h_1 < \ldots < h_S = r_0$ ranging from the smallest scale $h_0 = 0$ to the largest scale $h_S = r_0$ at each $d \in \mathcal{D}$ (Fig. 2(b)). By setting $\omega(d, d'; h_0) = 1$, we can estimate $\boldsymbol{\theta}(d)$ at scale h_0, which is denoted by $\hat{\boldsymbol{\theta}}(d; h_0)$, and construct a test statistic $W_\mu(d, h_0)$. Then, on the basis of the information that is 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 equations (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

$$\tilde{\omega}(d, d'; h) = \omega(d, d'; h) \Big/ \sum_{d' \in B(d,h)} \omega(d, d'; h).$$

For the sphere with radius h of the voxel d, on the basis of model (3), we consider a normalized weighted quasi-likelihood function $l_n\{\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}}\}$, which is given by

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

The $l_n\{\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}}\}$ 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 estimate of $\boldsymbol{\theta}(d)$, which is denoted by $\hat{\boldsymbol{\theta}}(d, h)$, is defined by

$$\hat{\boldsymbol{\theta}}(d, h) = \arg\max_{\boldsymbol{\theta}(d)} [n^{-1} l_n\{\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}}\}]. \tag{9}$$

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

$$\mathrm{cov}\{\hat{\boldsymbol{\theta}}(d, h)\} \approx \Sigma_n\{\hat{\boldsymbol{\theta}}(d, h)\} = \Sigma_{n,1}\{\hat{\boldsymbol{\theta}}(d, h)\}^{-1} \Sigma_{n,2}\{\hat{\boldsymbol{\theta}}(d, h)\} \Sigma_{n,1}\{\hat{\boldsymbol{\theta}}(d, h)\}^{-1}, \tag{10}$$

where $\Sigma_{n,1}\{\boldsymbol{\theta}(d)\} = -\partial^2_{\boldsymbol{\theta}(d)} l_n\{\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}}\}$ and

$$\Sigma_{n,2}\{\boldsymbol{\theta}(d)\} = \sum_{i=1}^{n} \left(\sum_{d' \in B(d,h)} \tilde{\omega}(d,d';h)\partial_{\boldsymbol{\theta}(d)} \log[p\{Y_i(d')|\mathbf{x}_i, \boldsymbol{\theta}(d)\}] \right)^{\otimes 2},$$

in which $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ 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:

$$H_{0,\mu} : R\{\boldsymbol{\theta}(d)\} = \mathbf{b}_0 \quad \textit{versus} \quad H_{1,\mu} : R\{\boldsymbol{\theta}(d)\} \neq \mathbf{b}_0, \tag{11}$$

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

$$W_\mu(d,h) = (R\{\hat{\boldsymbol{\theta}}(d;h)\} - \mathbf{b}_0)^{\mathrm{T}} \{\partial_{\boldsymbol{\theta}(d)} R\{\hat{\boldsymbol{\theta}}(d;h)\} \Sigma_n\{\hat{\boldsymbol{\theta}}(d;h)\} \partial_{\boldsymbol{\theta}(d)} R\{\hat{\boldsymbol{\theta}}(d;h)\}^{\mathrm{T}}\}^{-1} [R\{\hat{\boldsymbol{\theta}}(d;h)\} - \mathbf{b}_0]. \tag{12}$$

A path diagram of the MAET procedure is given below:

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

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

The MAET procedure consists of five key steps:

(a) initialization,
(b) weights adaptation,
(c) estimation,
(d) stop checking and
(e) inference.

In the initialization step (a), 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 neighbouring voxels and thus it can prevent oversmoothing $\boldsymbol{\theta}(d)$ at the beginning of the MAET procedure, 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 maximum weighted quasi-likelihood estimate $\hat{\boldsymbol{\theta}}(d,h_0)$, which is defined in equation (9) at each voxel $d \in \mathcal{D}$. The $\hat{\boldsymbol{\theta}}(d,h_0)$ are the same as those from the voxelwise approach. We then set $s = 1$ and $h_1 = c_h$.

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

$$D_{\boldsymbol{\theta}}(d,d';h_{s-1}) = (\hat{\boldsymbol{\theta}}(d,h_{s-1}) - \hat{\boldsymbol{\theta}}(d',h_{s-1}))^{\mathrm{T}} \Sigma_n\{\hat{\boldsymbol{\theta}}(d;h_{s-1})\}^{-1} (\hat{\boldsymbol{\theta}}(d,h_{s-1}) - \hat{\boldsymbol{\theta}}(d',h_{s-1})), \tag{14}$$

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

where $K_{\mathrm{loc}}(u)$ and $K_{\mathrm{st}}(u)$ are two non-negative kernel functions with compact support such that all of them decrease to 0 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 $K_{\mathrm{loc}}(\|d-d'\|_2/h_s)$ give less weight to the voxel $d' \in B(d, h_s)$, whose location is far from the voxel d. The weights $K_{\mathrm{st}}(u)$ 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 (c), for the radius h_s, we substitute $\omega(d, d'; h_s)$ into equation (9) to calculate $\hat{\boldsymbol{\theta}}(d, h_s)$ and then compute $W_\mu(d, h_s)$ according to equation (12) at each voxel $d \in \mathcal{D}$.

In the stop checking step (d), after the S_0th 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

$$D\{\hat{\boldsymbol{\theta}}(d; h_{S_0}), \hat{\boldsymbol{\theta}}(d; h_s)\} = (\hat{\boldsymbol{\theta}}(d, h_{S_0}) - \hat{\boldsymbol{\theta}}(d, h_s))^{\mathrm{T}} \Sigma_n\{\hat{\boldsymbol{\theta}}(d; h_{S_0})\}^{-1} (\hat{\boldsymbol{\theta}}(d, h_{S_0}) - \hat{\boldsymbol{\theta}}(d, h_s)). \quad (16)$$

Then, we check whether $\hat{\boldsymbol{\theta}}(d; h_s)$ is in an α confidence ellipsoid of $\hat{\boldsymbol{\theta}}(d; h_{S_0})$ given by $\{\boldsymbol{\theta}: D\{\hat{\boldsymbol{\theta}}(d; h_{S_0}), \boldsymbol{\theta}(d)\} \leqslant \tilde{C} = \chi^2(p)^\alpha\}$, 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 $D\{\hat{\boldsymbol{\theta}}(d; h_{S_0}), \hat{\boldsymbol{\theta}}(d; h_s)\} > \tilde{C}$, 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 (e). If $s \leqslant S_0$ or $D\{\hat{\boldsymbol{\theta}}(d; h_{S_0}), \hat{\boldsymbol{\theta}}(d; h_s)\} \leqslant \tilde{C}$ for $S - 1 \geqslant s > S_0$, then we set $h_{s+1} = c_h h_s$, increase s by 1 and continue with the weight adaptation step (b).

In the inference step (e), 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 method (Benjamini and Hochberg, 1995) or random-field theory (Worsley *et al.*, 2004; Nichols and Hayasaka, 2003), and then stop the algorithm.

2.3.1. Example 4

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

$$\hat{\boldsymbol{\beta}}(d, h_0)^{(t+1)} = \left[\sum_{i=1}^n X_i^{\mathrm{T}} \{\hat{\Sigma}(d, h_0)^{(t)}\}^{-1} X_i\right]^{-1} \sum_{i=1}^n X_i^{\mathrm{T}} \{\hat{\Sigma}(d, h_0)^{(t)}\}^{-1} Y_i(d), \quad (17)$$

$$\hat{\Sigma}(d, h_0)^{(t+1)} = (n - p_1)^{-1} \sum_{i=1}^n \{Y_i(d) - X_i \hat{\boldsymbol{\beta}}(d, h_0)^{(t+1)}\}^{\otimes 2}$$

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

$$\mathrm{cov}\{\hat{\boldsymbol{\beta}}(d, h_0)\} \approx \Sigma_n\{\hat{\boldsymbol{\beta}}(d, h_0)\} = \left\{\sum_{i=1}^n X_i^{\mathrm{T}} \hat{\Sigma}(d, h_0)^{-1} X_i\right\}^{-1}.$$

In the weight adaption step (b), compute

$$D_\beta(d, d'; h_{s-1}) = (\hat{\boldsymbol{\beta}}(d, h_{s-1}) - \hat{\boldsymbol{\beta}}(d', h_{s-1}))^{\mathrm{T}} \Sigma_n\{\hat{\boldsymbol{\beta}}(d; h_{s-1})\}^{-1} \{\hat{\boldsymbol{\beta}}(d, h_{s-1}) - \hat{\boldsymbol{\beta}}(d', h_{s-1})\}$$

and

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

In the estimation step (c), for the radius h_s, let

$$A(d, h_s, \boldsymbol{\omega}; X) = \sum_{i=1}^n X_i^{\mathrm{T}} \sum_{d' \in B(d, h_s)} \omega(d, d'; h_s)\, \hat{\Sigma}(d', h_0)^{-1} X_i,$$

compute

$$\hat{\boldsymbol{\beta}}(d,h_s)=A(d,h_s,\boldsymbol{\omega};X)^{-1}\left\{\sum_{i=1}^{n}X_i^{\mathrm{T}}\sum_{d'\in B(d,h_s)}\omega(d,d';h_s)\hat{\Sigma}(d',h_0)^{-1}Y_i(d')\right\}$$

and

$$\Sigma_n\{\hat{\boldsymbol{\beta}}(d,h_s)\}=A(d,h_s,\boldsymbol{\omega};X)^{-1}\left\{\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{\varepsilon}_i(d;\boldsymbol{\omega},h_s)^{\otimes 2}X_i\right\}A(d,h_s,\boldsymbol{\omega};X)^{-1},$$

where

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

In the stop checking step (d), we compute

$$D\{\hat{\boldsymbol{\beta}}(d;h_{S_0}),\hat{\boldsymbol{\beta}}(d;h_s)\}=(\hat{\boldsymbol{\beta}}(d,h_{S_0})-\hat{\boldsymbol{\beta}}(d,h_s))^{\mathrm{T}}\Sigma_n\{\hat{\boldsymbol{\beta}}(d;h_{S_0})\}^{-1}(\hat{\boldsymbol{\beta}}(d,h_{S_0})-\hat{\boldsymbol{\beta}}(d,h_s))$$

for $s>S_0$.

2.4. Parameters of the multiscale adaptive estimation testing procedure

The performance of the MAET procedure depends on specifying its following parameters: c_h, C_n, $K_{\mathrm{loc}}(u)$, $K_{\mathrm{st}}(u)$, S_0 and S. We have tested different combinations of these parameters of the MAET procedure in both simulated and real imaging data. According to our experience, the performance of the MAET procedure is quite robust to moderate changes to 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 or 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 includes only the closest neighbouring 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 the MAET procedure 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 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 the MAET procedure 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 the MAET procedure 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 the MAET procedure. 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 χ^2-distribution.

The $K_{\mathrm{loc}}(u)$ is a regular kernel function for further smoothing curves or surfaces based on the Euclidean distance between voxels. Some common choices of $K_{\mathrm{loc}}(u)$ include the Epanechnikov kernel (Tabelow *et al.*, 2006, 2008a, b, c; Polzehl and Spokoiny, 2000, 2006). Because the MAET

procedure mainly uses the similarity information between any pairs of voxels, the specification of $K_{\mathrm{loc}}(u)$ is not critical for it. We use $K_{\mathrm{loc}}(u) = (1-u)_+$.

We set $K_{\mathrm{st}}(u) = \exp(-u)$ in our simulated and real imaging data. Theoretically, as shown later, $\exp(-u)$ gives an exponential decay rate of n. Although different choices of $K_{\mathrm{st}}(\cdot)$ 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 $K_{\mathrm{st}}(u) = \exp(-u)$ performs reasonably well. Another good choice of $K_{\mathrm{st}}(u)$ is $\min(1, 2(1-u))_+$, which has better performance in spatially and adaptively smoothing FMRI and diffusion tensor images 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_0th iteration, the stop checking step starts to compute the stopping criterion and to 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 to prevent bad data from neighbouring 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 neighbouring 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 down-weights voxel $d' \in B(d, h)$ in $l_n\{\boldsymbol{\theta}(d); h, \tilde{\omega}\}$ when the value of $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$ 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 $\hat{\boldsymbol{\theta}}(d, h_s)^{(0)} = \hat{\boldsymbol{\theta}}(d, h_{s-1})$, the Newton–Raphson algorithm converges very fast. The additional computational time for the MARM is moderate compared with the voxelwise approach, since the MARM involves only 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 the MAET procedure with stochastic adaptive weights. A critical question is what kinds of stochastic weights can automatically incorporate 'good' information and prevent 'bad' information from neighbouring voxels? By appropriately utilizing information from neighbouring 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 that are 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 behaviour of the adaptive weight when $s = 1$ and then discuss the scenario when $s > 1$.

Throughout the paper, we consider only 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 the MAET procedure, 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 data set is always fixed. We obtain the following theorems, whose detailed assumptions and proofs can be found in the supplementary report Li *et al.* (2010).

Theorem 1. If assumptions (C1)–(C7) in Li *et al.* (2010) are true, then we have

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

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

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 $\Sigma_{1*}(d) = -E(\partial^2_{\boldsymbol{\theta}(d)} \log[p\{Y(d)|\mathbf{x}, \boldsymbol{\theta}_*(d)\}])$ and $\Sigma_{2*}(d) = E\{(\partial_{\boldsymbol{\theta}(d)} \log[p\{Y(d)|\mathbf{x}, \boldsymbol{\theta}_*(d)\}])^{\otimes 2}\}$, and

(d) for any $\varepsilon_0 > 0$, $\lim_{n\to\infty}(P[|K_{\mathrm{st}}\{D_{\boldsymbol{\theta}}(d,d';h_0)/C_n\} - \mathbf{1}\{\Delta_*(d,d')=\mathbf{0}\}| > \varepsilon_0]) = 0$.

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

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

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

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

Theorem 2 shows that the MAET procedure has several remarkable features. Theorem 2, part (a), ensures that $\hat{\boldsymbol{\theta}}(d,h)$ is a consistent estimate of $\boldsymbol{\theta}_*(d)$ for the adaptive weights in equation (15) for any $h > 0$. Theorem 2, part (b), ensures that $\hat{\boldsymbol{\theta}}(d,h)$ is a root n estimate of $\boldsymbol{\theta}_*(d)$. Theorem 2, part (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 α-percentile of the $F_{r,n-r}$-distribution, i.e. we reject H_0 if $W_\mu(d,h) \geqslant r(n-1)F^{1-\alpha}_{r,n-r}/(n-r)$, and do not reject H_0 otherwise.

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

Corollary 1. Suppose that 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, part (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, part (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 set-up. Moreover, if there is a voxel $d'\in B(d,h_1)/\{d\}$ such that $\boldsymbol{\theta}_*(d)=\boldsymbol{\theta}_*(d')$, corollary 1, part (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 $Y_i(d)=\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(d)+\varepsilon_i(d)$, where $\varepsilon_i(d)\sim N\{0,\tau(d)^{-1}\}$. Let

$$\tilde{\omega}_\tau(d,d';h)=\tau(d')\omega(d,d';h)\Big/\sum_{d'\in B(d,h)}\tau(d')\omega(d,d';h);$$

we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}(d,h)&=\left(\sum_{i=1}^n\mathbf{x}_i^{\otimes 2}\right)^{-1}\sum_{i=1}^n\mathbf{x}_i\,Y_i(d;\tilde{\boldsymbol{\omega}}_\tau,h),\\ \operatorname{cov}\{\hat{\boldsymbol{\beta}}(d,h)\}&\approx\left(\sum_{i=1}^n\mathbf{x}_i^{\otimes 2}\right)^{-1}\sum_{i=1}^n\mathbf{x}_i^{\otimes 2}\hat{\varepsilon}_i(d;\tilde{\boldsymbol{\omega}}_\tau,h)^2\left(\sum_{i=1}^n\mathbf{x}_i^{\otimes 2}\right)^{-1},\end{aligned}\tag{19}$$

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

$$\hat{\varepsilon}_i(d;\tilde{\boldsymbol{\omega}}_\tau,h)=\sum_{d'\in B(d,h)}\tilde{\omega}_\tau(d,d';h)\{Y_i(d')-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(d',h)\}.$$

Although Tabelow *et al.* (2006) have obtained the same $\hat{\boldsymbol{\beta}}(d,h)$ as in expression (19), the MARM that is developed here has several advantages. We shall 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 expression (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 $n^{1/2}\{\hat{\boldsymbol{\beta}}(d,h)-\boldsymbol{\beta}_*\}$ is asymptotically equivalent to

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

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

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

where $\{Z(d'):d'\in B(d,h)\}$ is a Gaussian vector with mean 0 and covariance structure $\mathrm{cov}\{Z(d)\}=\tau(d)^{-1}\mathbf{I}_{p_1}$ and $\mathrm{cov}\{Z(d),Z(d')\}=E\{\varepsilon_1(d)\,\varepsilon_1(d')\}\mathbf{I}_{p_1}$.

(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 $n^{1/2}\{\hat{\beta}(d,h_1)-\beta_*\}$ is asymptotically equivalent to

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

where

$$\mathcal{E}_n(d,d')=\tau(d)\,\mathrm{tr}([E(\mathbf{x}^{\otimes 2})^{-1/2}n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\{\varepsilon_i(d)-\varepsilon_i(d')\}]^{\otimes 2}).$$

As $n\to\infty$, $A_2(d;h_1)$ converges in distribution to a random vector given by

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

Theorem 3 gives a theoretical justification of the multiscale adaptive linear model. Theorem 3, parts (a) and (b), formally characterize the key differences between a bounded and unbounded C_n in the linear model. Theorem 3, part (a), shows that, for certain unbounded C_n, the asymptotic distributions of $\hat{\beta}(d,h)$ are always normally distributed. For a bounded C_n, however, theorem 3, part (b), only gives the asymptotic distribution of $\hat{\beta}(d,h_1)$, 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 $\hat{\beta}(d,h)$ and $W_\mu(d,h)$ with respect to different scales h and compare the MARM with the voxelwise method. For brevity, we only present some results based on a 64×64 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 $Y_i(d)=\mathbf{x}_i^{\mathrm{T}}\beta(d)+\varepsilon_i(d)$ for $i=1,\ldots,n$, where $\beta(d)=(\beta_1(d),\beta_2(d),\beta_3(d))^{\mathrm{T}}$ and $\mathbf{x}_i=(1,x_{i2},x_{i3})^{\mathrm{T}}$. Errors $\varepsilon_i(d)$ were first independently generated from $N(0,1)$ and $\chi^2(3)-3$ distributions and then they were smoothed by using heat kernel smoothing with four iterations, which gave an effective smoothness of about two 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 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

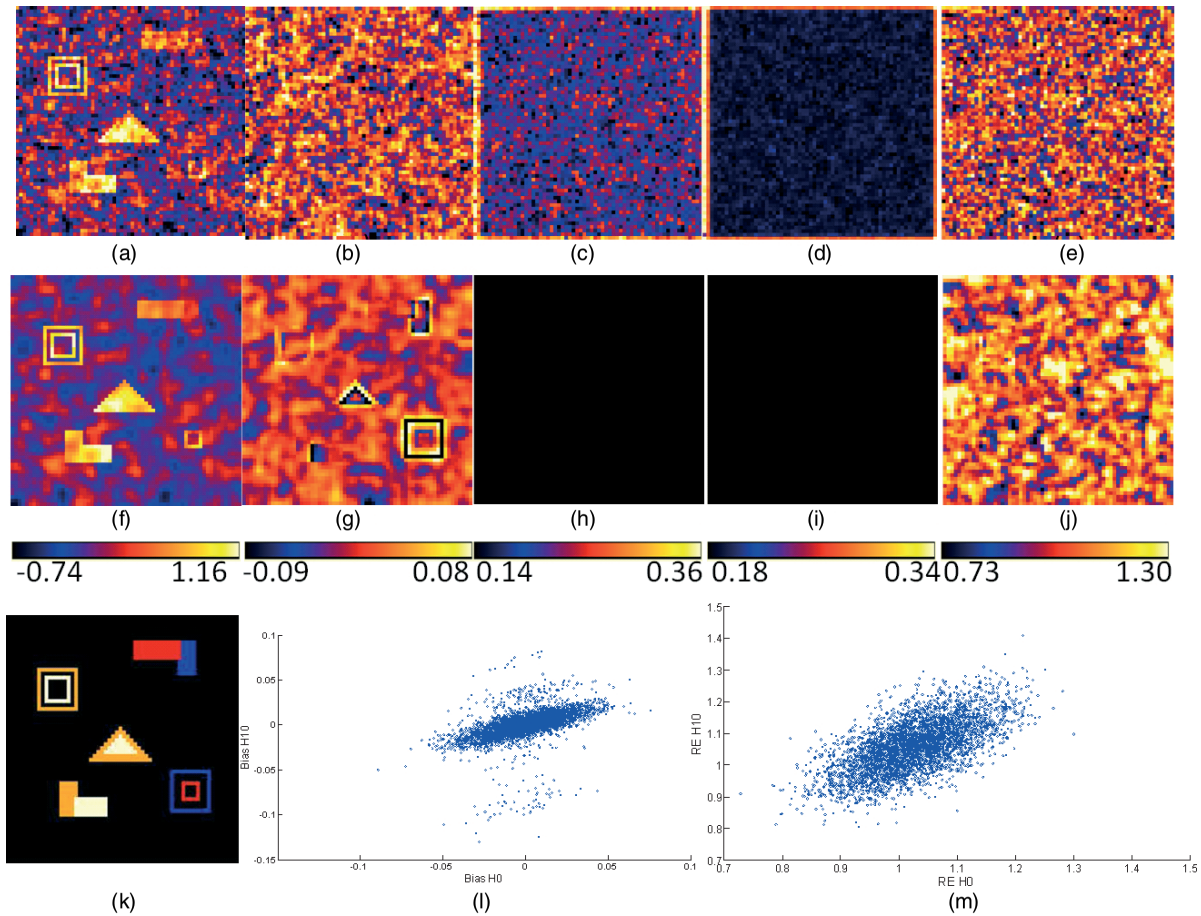


Fig. 3. Results from a simulation study of comparing (a)–(e) the voxelwise method and (f)–(j) the MARM (MAET procedure, $S = 10$ and $c_n = 1.1$) on the basis of 1000 $N(0,1)$-distributed data with $n = 60$: (a) selected image of $\hat{\beta}_2(d, h_0)$ obtained from a simulated data set; (b) bias image of $\hat{\beta}_2(d, h_0)$; (c) RMS-image of $\hat{\beta}_2(d, h_0)$; (d) SD-image of $\hat{\beta}_2(d, h_0)$; (e) RE-image of $\hat{\beta}_2(d, h_0)$; (f) 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 $\hat{\beta}_2(d, h_{10})$; (j) RE-image of $\hat{\beta}_2(d, h_{10})$; (k) ground truth image of five ROIs with black, blue, red, yellow and white representing $\beta_2(d) = 0$, $\beta_2(d) = 0.2$, $\beta_2(d) = 0.4$, $\beta_2(d) = 0.6$ and $\beta_2(d) = 0.8$ respectively; (l) scatter plot of biases of $\hat{\beta}_2(d, h_0)$ *versus* $\hat{\beta}_2(d, h_{10})$; (m) scatter plots of REs of $\hat{\beta}_2(d, h_0)$ *versus* $\hat{\beta}_2(d, h_0)$

$\beta_1(d) = \beta_3(d) = 0$ 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 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 the MARM can perform well for different shapes. The true $\beta_2(d)$ was displayed for all ROIs with black, blue, red, yellow and white colours representing $\beta_2(d)$ = 0, 0.2, 0.4, 0.6, 0.8 (Fig. 3(k)).

Table 1. Average bias ($\times 10^{-3}$), RMS, SD, RE and MVR of $\beta_2(d)$ parameters in the five ROIs at three different scales (h_0, h_5, h_{10}), two different distributions ($N(0, 1)$ and $\chi^2(3) - 3$ distributions) and two different sample sizes ($n = 60, 80$)†

Parameter	*Results for* $\chi^2(3) - 3$						*Results for* $N(0, 1)$					
	$n = 60$			$n = 80$			$n = 60$			$n = 80$		
	h_0	h_5	h_{10}	h_0	h_5	h_{10}	h_0	h_5	h_{10}	h_0	h_5	h_{10}
$\beta_2(d) = 0.0$												
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
MVR	1.00	0.59	0.44	1.00	0.61	0.46	1.00	0.63	0.46	1.00	0.64	0.47
$\beta_2(d) = 0.2$												
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
MVR	1.00	0.70	0.50	1.00	0.71	0.51	1.00	0.72	0.52	1.00	0.73	0.52
$\beta_2(d) = 0.4$												
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
MVR	1.00	0.70	0.50	1.00	0.70	0.51	1.00	0.71	0.52	1.00	0.72	0.52
$\beta_2(d) = 0.6$												
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
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
$\beta_2(d) = 0.8$												
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

†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, 1000 simulated data sets were used.

We fitted the linear model $Y_i(d) = \mathbf{x}_i^{\mathrm{T}}\beta(d) + \varepsilon_i(d)$, where $\varepsilon_i(d) \sim N\{0, \tau(d)^{-1}\}$, and then applied the MAET procedure that was 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, which is 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 1000 simulated data sets. For brevity, we present only the results for $\hat{\beta}_2(d, h_0)$ and $\hat{\beta}_2(d, h_{10})$ obtained from $N(0, 1)$-distributed data with $n = 60$ in Fig. 3. We also calculated the average bias, RMS, SD, RE and maximum VR, MVR, in each of the five ROIs and present them in Table 1. The biases are slightly increased from h_0 to h_{10} (Figs 3(b) and 3(g) and Table 1), whereas RMS and SD at h_5 and h_{10} are much smaller than those at h_0 (Figs 3(c), 3(d), 3(h) and 3(i) and Table 1). In addition, RMS and its corresponding SD are relatively close to each other at all scales for both the normally and the χ^2-distributed data (Table 1 and Figs 3(e) and 3(j)). Moreover, the average SDs and MVRs in ROIs with $\beta_2(d) > 0$ are larger than those in ROIs with $\beta_2(d) = 0$ (Figs 3(i) and Table 1), because the interior of the ROI with $\beta_2(d) = 0$ contains more pixels (Fig. 3(k)). The biases, SDs and RMSs of $\beta_2(d)$ are smaller in the normally distributed data than in the χ^2-distributed data (Table 1), because the signal-to-noise ratios in the normally distributed data are 2.45 times bigger than the signal-to-noise ratios in the χ^2-distributed data. Increasing the sample size and signal-to-noise ratio decreases the bias, RMS and SD 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 type 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 1000 replications were used to calculate the estimates and standard errors of rates of rejection at $\alpha = 5\%$ significance level. For $W_\mu(d, h)$, the type I error rates in the ROI with $\beta_2(d) = 0$ were relatively accurate for all scales, whereas the statistical power for rejecting the null hypothesis in ROIs with $\beta_2(d) \neq 0$ was significantly increased with radius h and signal-to-noise ratio (Table 2).

Table 2. Simulation study for $W_\mu(d, h)$†

$\beta_2(d)$	h_s	*Results for N(0, 1)*				*Results for* $\chi^2(3) - 3$			
		$n = 60$		$n = 80$		$n = 60$		$n = 80$	
		Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
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

†Estimates and standard errors of rates of rejection for pixels inside the five ROIs were reported at two different scales (h_0, h_{10}), two different distributions ($N(0, 1)$ and $\chi^2(3) - 3$) and two different sample sizes ($n = 60$ and $n = 80$) at $\alpha = 5\%$. For each case, 1000 simulated data sets were used.

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 is its ability to reveal the white matter maturation process in the human brain by using a set of water-diffusion-related parameters, such as FA and radial diffusivity. For instance, FA is a measure representing the inhomogeneous extent of local barriers to water diffusion. FA has been widely used to investigate early brain development from identifying transient brain structures such as ganglionic eminence and cortical subpliate to estimating the correlation of white matter maturation with functional development measures such as intelligence and working memory.

We considered 38 subjects from the neonatal project on early brain development that was 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. The 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 mm^{-2} and a $b=0$ reference scan. 46 contiguous slices with a slice thickness of 2 mm covered a field of view of 256×256 mm^2 with an isotropic voxel size of $2\times2\times2$ mm^3. High resolution T1-weighted images were acquired by using a three-dimensional 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 diffusion tensor images (38 subjects, three time points each) were registered to a randomly selected brain diffusion tensor image of a 2-year-old subject by using tensor image morphing for elastic registration (Yap *et al.*, 2009).

FA calculated from diffusion tensor images 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 $i=1,\ldots,38$ and $j=1,2,3$, at each voxel of the template, where t_{ij} denotes the jth scan time for the ith subject, $\varepsilon_i(d)=(\varepsilon_{i1}(d),\varepsilon_{i2}(d),\varepsilon_{i3}(d))^{\mathrm{T}}\sim N\{\mathbf{0},\Sigma(d)\}$ and $\Sigma(d)$ is a 3×3 unstructured covariance matrix. The MAET procedure that was 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 by using the Bonferroni correction with overall level of significance 1%. As S increases from 0 to 10, the MARM shows a clear advantage in detecting more significant and smoothed significant areas as well as preserving the edges of grey matter, white matter and cerebrospinal fluid areas (Figs 1(a)–1(d) and 1(h)). We also smoothed FA imaging data by using an isotropic Gaussian kernel with full width at half maximum 6 mm and then analysed the data by using the voxelwise approach. The results based on the smoothed FA images show the obvious oversmoothing in cerebrospinal fluid and the grey matter areas, such as the ventricle (Figs 1(a)–1(c)). Furthermore, we identified a voxel in the red circle in the ventricle, whose location is near the boundary of the white matter and cerebrospinal fluid (see the red circle in Fig. 1(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 Fig. 1(a) does not reveal any growth patterns, which agrees with the fact that the ventricle contains cerebrospinal fluid in the brain (Fig. 1(d)). However, after being smoothed with the Gaussian kernel, smoothed FA values gradually increase with age (Fig. 1(h)). This indicates that the data in the red voxel were oversmoothed because its neighbouring voxels contain 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 1(e)–1(g)). Major white matter structures

are already presented in FA at birth (Fig. 1(e)). Within the central brain region, different developing patterns were observed for the *genu*, *splenium* and body of *corpus callosum*, internal and external capsules (Figs 1(i)–1(l)). In FA, the *genu* and *splenium* have a similar FA value at birth and the *genu*'s FA gradually increases higher than the *splenium*'s. The *corpus callosum* body has a slightly lower FA compared with 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 paper studies the idea of using an MARM for the spatial and adaptive analysis of neuroimaging data. The MARM integrates the PS approach with the voxelwise method for neuroimaging data from multiple subjects. There are three features in the MARM: being spatial, being hierarchical and being adaptive. The 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 between different voxels. Thus, the MARM explicitly utilizes the spatial information to carry out statistical inference. The MARM also builds hierarchically nested spheres by increasing the radius of a spherical neighbourhood around each voxel and utilizes information in each of the nested spheres across all voxels. Finally, the MARM combines all observations with adaptive weights in the voxels within the sphere of the current voxel to calculate parameter estimates and test statistics adaptively. Without imposing any spatial correlation patterns, we have derived the asymptotic properties of the parameter estimates and test statistics for the 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 appropriate values of various parameters in the MAET procedure.

Many issues still merit further research. The three key features of the 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 non-parametric components. More research is needed for optimizing the choices of parameters in the MAET procedure and weakening 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 is to develop adaptive neighbourhood methods to determine multiscale neighbourhoods that adapt to the pattern of imaging data at each voxel. An important issue is that the voxelwise approach and the MARM are also based on the perfect registration assumption, which is demonstrably false. We may need to integrate the registration method, smoothing method and voxelwise approach into a unified framework so that we can appropriately account for registration errors in the statistical analysis.

Acknowledgements

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

References

Ashburner, J. and Friston, K. J. (2000) Voxel-based morphometry: the methods. *NeuroImage*, **11**, 805–821.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall.

Basser, P. J., Mattiello, J. and LeBihan, D. (1994) MR diffusion tensor spectroscopy and imaging. *Biophys. J.*, **66**, 259–267.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *J. R. Statist. Soc.* B, **48**, 259–302.

Bowman, F. D. (2007) Spatio-temporal models for region of interest analyses of functional mapping experiments. *J. Am. Statist. Ass.*, **102**, 442–453.

Chung, M. K., Robbins, S., Dalton, K. M., Davidson, R. J., Alexander, A. L. and Evans, A. C. (2005) Cortical thickness analysis in autism via heat kernel smoothing. *NeuroImage*, **25**, 1256–1265.

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

Hecke, W. V., Sijbers, J., Backer, S. D., Poot, D., Parizel, P. M. and Leemans, A. (2009). On the construction of a ground truth framework for evaluating voxel-based diffusion tensor MRI analysis methods. *NeuroImage*, **46**, 692–707.

Huettel, S. A., Song, A. W. and McCarthy, G. (2004) *Functional Magnetic Resonance Imaging*. Sunderland: Sinauer Associates.

Jones, D. K., Symms, D. K., Cercignani, M. and Howard, R. J. (2005) The effect of filter size on VBM analyses of DT-MRI data. *NeuroImage*, **26**, 546–554.

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

Li, Y. M., Zhu, H., Shen, D., Lin, W., Gilmore, J. H. and Ibrahim, J. G. (2010) Technical proofs for Multiscale adaptive regression models for neuroimaging data. Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill. (Available from `http://www.bios.unc.edu/research/bias`.)

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Nichols, T. and Hayasaka, S. (2003) Controlling the family-wise error rate in functional neuroimaging: a comparative review. *Statist. Meth. Med. Res.*, **12**, 419–446.

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

Polzehl, J. and Spokoiny, V. G. (2006) Propagation-separation approach for local likelihood estimation. *Probab. Theor. Reltd Flds*, **135**, 335–362.

Polzehl, J. and Tabelow, K. (2007) fmri: a package for analyzing fmri data. *R News*, **7**, 13–17.

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

Qiu, P. (2007) Jump surface estimation, edge detection, and image restoration. *J. Am. Statist. Ass.*, **102**, 745–756.

Tabelow, K., Piech, V., Polzehl, J. and Voss, H. U. (2008a) High-resolution fMRI: overcoming the signal-to-noise problem. *J. Neursci. Meth.*, **178**, 357–365.

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

Tabelow, K., Polzehl, J., Ulug, A. M., Dyke, J. P., Watts, R., Heier, L. A. and Voss, H. U. (2008c) Accurate localization of functional brain activity using structure adaptive smoothing. *IEEE Trans. Med. Imgng*, **27**, 531–537.

Tabelow, K., Polzehl, J., Voss, H. U. and Spokoiny, V. (2006) Analyzing fMRI experiments with structural adaptive smoothing procedures. *NeuroImage*, **33**, 55–62.

Thompson, P. M. and Toga, A. W. (2002) A framework for computational anatomy. *Comput. Visualizn Sci.*, **5**, 13–34.

Worsley, K. J. (2003) Detecting activation in fMRI data. *Statist. Meth. Med. Res.*, **12**, 401–418.

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

Yap, P. T., Wu, G. R., Zhu, H.T., Lin, W. L. and Shen, D. G. (2009) TIMER: tensor image morphing for elastic registration. *NeuroImage*, **47**, 549–563.

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

Zhu, H. T., Gu, M. G. and Peterson, B. G. (2007a) Maximum likelihood from spatial random effects models via the stochastic approximation expectation maximization algorithm. *Statist. Comput.*, **15**, 163–177.

Zhu, H. T., Zhang, H. P., Ibrahim, J. G. and Peterson, B. S. (2007b) Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance image data (with discussion). *J. Am. Statist. Ass.*, **102**, 1085–1102.