Statistical Analysis of Complex Neuroimaging Data

Yimei Li

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2009

Approved by:

Hongtu Zhu, Advisor

Joseph G. Ibrahim, Advisor

Jianwen Cai, Reader

John Gilmore, Reader

Dinggang Shen, Reader

Donglin Zeng, Reader

Abstract

YIMEI LI: Statistical Analysis of Complex Neuroimaging Data.
(Under the direction of Hongtu Zhu and Joseph G. Ibrahim.)

This dissertation is composed of two major topics: a) regression models for identifying noise sources in magnetic resonance images, and b) multiscale Adaptive method in neuroimaging studies.

The first topic is covered by the first thesis paper. In this paper, we formally introduce three regression models including a Rician regression model and two associated normal models to characterize stochastic noise in various magnetic resonance imaging modalities, including diffusion weighted imaging (DWI) and functional MRI (fMRI). Estimation algorithms are introduced to maximize the likelihood function of the three regression models. We also develop a diagnostic procedure for systematically exploring MR images to identify noise components other than simple stochastic noise, and to detect discrepancies between the fitted regression models and MRI data. The diagnostic procedure includes goodness-of-fit statistics, measures of influence, and tools for graphical display. The goodness-of-fit statistics can assess the key assumptions of the three regression models, whereas measures of influence can isolate outliers caused by certain noise components, including motion artifact. The tools for graphical display permit graphical visualization of the values for the goodness-of-fit statistic and influence measures. Finally, we conduct simulation studies to evaluate performance of these methods, and we analyze a real dataset to illustrate how our diagnostic procedure localizes subtle image artifacts by detecting intravoxel variability that is not captured by the regression models.

The second topic, multiscale adaptive methods for neuroimaging data, consists of two

thesis papers.The goal of the first paper is to develop a multiscale adaptive regression model (MARM) for spatial and adaptive analysis of neuroimaging data. Compared with the existing voxel-wise approach in the analysis of imaging data,MARM has three unique features: being spatial, being hierarchical, and being adaptive. MARM creates a small sphere with a given radius at each location (called voxel), analyzes all observations in the sphere of each voxel, and then uses these consecutively connected spheres across all voxels to capture spatial dependence among imaging observations. MARM builds hierarchically nested spheres by increasing the radius of a spherical neighborhood around each voxel and utilizes information in each of the nested spheres at each voxel. Finally, MARM combine imaging observations with adaptive weights in the voxels within the sphere of the current voxel to adaptively calculate parameter estimates and test statistics. Theoretically, we establish the consistency and asymptotic normality of adaptive estimates and the asymptotic distributions of adaptive test statistics under some mild conditions. Three sets of simulation studies are used to demonstrate the methodology and examine the finite sample performance of the adaptive estimates and test statistics in MARM. We apply MARM to quantify spatiotemporal white matter maturation patterns in early postnatal population using diffusion tensor imaging. Our simulation studies and real data analysis confirm that the MARM significantly outperforms the voxel-wise methods.

The goal of the second paper is to develop a multiscale adaptive generalized estimation equation (MAGEE) for spatial and adaptive analysis of longitudinal neuroimaging data. Longitudinal imaging studies have been valuable for better understanding disease progression and normal brain development/aging. Compared to cross-sectional imaging studies, longitudinal imaging studies can increase the statistical power in detecting subtle spatiotemporal changes of brain structure and function. MAGEE is a hierarchical, spatial, semiparametric, and adaptive procedure, compared with the existing voxel-wise approach. The key ideas of MAGEE are to build hierarchically nested spheres with

increasing radii at each location, to analyze all observations in the sphere of each voxel using weighted generalized estimating equations, and to use the consecutively connected spheres across all voxels to adaptively capture spatial pattern. Simulation studies and real data analysis clearly show the advantage of MAGEE method over the existing voxel-wise methods. Our results also reveal i) the increase of fractional anisotropy in this early postnatal stage, and ii) five different growth patterns in the brain regions under examination.

Acknowledgments

First, my most earnest acknowledgment goes to my advisor Dr. Hongtu Zhu for his mentorship, encouragement, inspiration, and support during the preparation of this dissertation. His enthusiasm for science and persistence for research set a great example for me to follow in my future career. Also, I would like to convey my appreciation to my co-advisor Dr. Joseph Ibrahim for his help, great lectures, and constant encouragement. I would also like to thank Dr. Donglin Zeng and Dr. Jianwen Cai for their help and comments. Many warm thanks also go to Dr. Dinggang Shen and Dr. John Gilmore for their important contributions and the kind offer for me to use the valuable datasets from their labs.

Finally, my final and most heartfelt, acknowledgment must go to my parents, Mingsheng Li and Guilan Zhang. They have been the source for the endless support, comfort and love for my whole life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADC	Apprent diffusion coefficient
CAR	Conditional autoregressive
CK	Conditinal Kolmogrov test
CM	Cramer-von Mises test
DTI	Diffusion tensor images
DWI	Diffusion weighted imaging
EPI	Echo-planar imaging
FDR	False discovery rate
fMRI	Functinal magnetic resounance imaging
FSL	FMRIB Software Library
GEE	Generalized estimating equation
IC	Independent component
ICA	Independent component analysis
LM	General linear model
LMM	Linear mixed effect model
MAGEE	Multiscale adaptive generalized estimation equation
MARM	Multiscale adptive regression model

MRF Markov random field

MRI Magnetic resounance imaging

MR Magnetic resounance

MWQL Maximum weighted quasi-likelihood

PS Propagation-separation

RFT Random field theory

ROI Region-of-interest

SNR Signal to noise ratio

SPM Statistical parametric mapping

Chapter 1

Introduction and literature review

This dissertation is composed of two major topics in imaging data analysis: First, regression models for identifying noise sources in magnetic resonance images. Second, multiscale Adaptive method in neuroimaging studies.

The first topic is covered by the first thesis paper. In this paper, we formally introduce three regression models including a Rician regression model and two associated normal models to characterize stochastic noise in various magnetic resonance imaging modalities, including diffusion weighted imaging (DWI) and functional MRI (fMRI). Estimation algorithms are introduced to maximize the likelihood function of the three regression models. We also develop a diagnostic procedure for systematically exploring MR images to identify noise components other than simple stochastic noise, and to detect discrepancies between the fitted regression models and MRI data. The diagnostic procedure includes goodness-of-fit statistics, measures of influence, and tools for graphical display. The goodness-of-fit statistics can assess the key assumptions of the three regression models, whereas measures of influence can isolate outliers caused by certain noise components, including motion artifact. The tools for graphical display permit graphical visualization of the values for the goodness-of-fit statistic and influence measures.

The second topic, multiscale adaptive methods for neuroimaging data, consists of

two thesis papers. The goal of the first paper is to develop a multiscale adaptive regression model (MARM) for spatial and adaptive analysis of neuroimaging data. Compared with the existing voxel-wise approach in the analysis of imaging data, MARM has three unique features: being spatial, being hierarchical, and being adaptive. MARM creates a small sphere with a given radius at each location (called voxel), analyzes all observations in the sphere of each voxel, and then uses these consecutively connected spheres across all voxels to capture spatial dependence among imaging observations. MARM builds hierarchically nested spheres by increasing the radius of a spherical neighborhood around each voxel and utilizes information in each of the nested spheres at each voxel. Finally, MARM combine imaging observations with adaptive weights in the voxels within the sphere of the current voxel to adaptively calculate parameter estimates and test statistics. Theoretically, we establish the consistency and asymptotic normality of adaptive estimates and the asymptotic distributions of adaptive test statistics under some mild conditions.

The goal of the second paper is to develop a multiscale adaptive generalized estimating equation (MAGEE) for the spatial and adaptive analysis of longitudinal neuroimaging data and to demonstrate its superiority over the voxel-wise approach using simulated and real imaging data. Compared with the Gaussian distributional assumption in the general linear model, MAGEE is a semiparametric method and explicitly account for the temporal correlation existed between the repeated measurements from the same subject. Thus, it is very desirable for the analysis of longitudinal neuroimaging data. MAGEE also includes specific methods for approximating the standard errors of the smoothed parametric estimates. We also theoretically examine the adaptive weights in the MAGEE and their roles in ensuring the proper statistical properties of parameter estimators. Finally, we formalize some technical conditions and formally establish the asymptotic properties including consistency and asymptotic distributions of the parameter estimates and test statistics for MAGEE.

The dissertation is organized as follows. The next section presents a literature review for each of the two topics. The first covers a review on diagnostic measures for missing data, and the second reviews existing statistical methods for neuroimaging studies. Then we proceed to present each of the three papers: We assess how to identify noise sources in magnetic resonance images by regression models in Chapter 2, and we formally develop multiscale adaptive regression models for neuroimaging data in Chapter 3 and multiscale adaptive generalized estimating equation (MAGEE) for the spatial and adaptive analysis of longitudinal neuroimaging data in Chapter 4.

1.1 Regression Models for Identifying Noise Sources in Magnetic Resonance Images

Magetnic resonance images contain various souces of temporal and spatial noises. The thermal motion of elctrons within the subject and within the scanner slectronics leads to the intrinsic thermal noise. The complicated imaging hardware system has its own error called system noise. In addtion to noises resulting from intrinsic properties of the magnetic resonance imaging, motion and phyisological noise is also one of the major sources of noise when human subjects are scanned through MRI system. For example, Muscle contraction, blood pulse, metabolism of neural system and large motions exists typically during MRI scanning (Huettel, Song, and McCarthy 2004). Previous studies have shown that those noise components can introduce substantial bias into measurements and estimation made from those images, such as indices for the principle direction of fiber tracts in diffusion tensor images (Skare, Li, Nordell, and Ingvar 2000; Luo and Nichols 2003; Nowark 1999). Correct understanding the noise components is essential for MRI data analysis.

The raw data obtained during MRI scanning are complex values that represent the Fourier transformation of a magnetization distribution of a volume of tissue at a certain point in time. An inverse Fourier transform converts these raw data into magnitude,

frequency, and phase components that more directly represent the physiological and morphological features of interest in the person being scanned. The magnetic susceptibility, chemical shift, and perfusion of tissues, for example, can be represented using either the magnitude or the phase angle of these Fourier-transformed data.

The electronic noise in the real and imaginary parts of the raw MR data are usually assumed to be independently Gaussian distributed (Henkelman 1985; Gudbjartsson and Patz 1995; Macovski 1996). Then, it can be shown theoretically that the Rician distribution is the model for characterizing the stochastic noise in the magnitude of MR data. Moreover, in practice, the Rician noise distribution of MR data has been experimentally validated using MR data (Haacke, Brown, Thompson, and Venkatesan 1999). Furthermore, the Rician distribution can be reasonably approximated by normal distributions at high signal-to-noise (SNR) ratios (Gudbjartsson and Patz 1995; Rowe and Logan 2005). Despite the extensive use of Rician and normal distributions in analyzing MR images (Kristoffersen 2007; Rowe 2005; Sijbers and den Dekker 2004; Sijbers, den Dekker, Scheunders, and van Dyck 1998a; Sijbers, den Dekker, Verhoye, van Audekerke, and van Dyck 1998b), a formal statistical framework for characterizing stochastic noise in various MR imaging modalities has not yet been developed. Rician regression model is needed for better understanding the noise components in MRI.

Other non-stochastic noise can cause the magnitude data of the MRI deviates from Rician distribution. Important tools to detect the outliers and influential oberservations in regression models are diagnostic measures. Residuals and Cook's distance have been widely used to identify influential observations in various regression models (Cox and Snell, 1968; Cook and Weisberg, 1982). Goodness-of-fit test statistics is used to identify the discrepancy between observed values and the values expected under the model in question. Influence measures based on case-deletion diagnostics have been studied extensively in regression models (Cook and Weisberg 1982; Wei 1998). However, the diagnostic tools for the Rician regression model have not been developed previously.

1.2 Multiscale Method for Neuroimaging Data

Magnetic resonance imaging becomes popular tool to study the detail and accurate mesures of brain morphology (Ashburner and Fristion, 2000; Chung, Robbins, Dalton, Davidson, Alexander and Evans, 2005; Styner, Lieberman,McClure, Weinberger, Jones and Gerig 2005, Thompson and Toga, 2002). There is an extensive literature on development of voxel-wise methods for analyzing high-dimensional data including particularly MRI measures on the 2D surface or the 3D volume. The existing voxel-wise methods for analyzing high-dimensional data are primarily executed in two sequential steps. The first step involves fitting a statistical model, such as general linear model (LM) and a linear mixed model (LMM), to data from all subjects at each location, such as voxel, and generating a statistical parametric map of test statistics (or p-values) (Friston et al., 1995; Beckmann, Jenkinson, and Smith, 2003). The second step is to compute adjusted p-values in order to account for testing multiple hypotheses across thousands to millions of locations using various statistical methods (e.g., random field theory (RFT), false discovery rate, or permutation methods) (Nichols and Hayasaka, 2003; Worsley et al., 2004).

The existing voxel-wise methods have some obvious limitations for the analysis of MRI imaging data, which underscore the great need for further methodological development. (i) In essence, the voxel-wise methods treats all voxels as independent units (Tabelow *et al.*, 2006). Neuroimaging data, however, are spatially correlated in nature and it is anticipated to observe spatially contiguous regions of activation with rather sharp edges in many neuroimaging studies. (ii) It is common to apply a smoothing step before applying for voxel-wise approach for analysis of neuroimaging data. Smoothing imaging data, however, blurs the image data near the edges of activated regions and thus it can dramatically increase the numbers of false positives and false negatives (Polzehl and Spokoiny, 2000, 2003, 2006; Qiu, 2005, 2007; Tabelow *et al.*, 2006). (iii) All voxel-wise approaches are also based on a stringent assumption that after an image warping

procedure, the location of a voxel in the image of one person is in precisely the same location as the voxel identified in another person, which is demonstrably false in neuroimaging data. (iv) More seriously, as a new imaging technique enables people to collect images with higher resolution, applying the voxel-wise methods to these new images, which contain much more voxels with smaller sizes, has much less statistical power in detecting statistically significant patterns. Besides high correlation, neuroimaging data in the neighboring voxels are strongly linked with each other and noisy homogeneous patches are usually expected.

Spatially modeling neuroimaging data in the 3D volume (or 2D surface) represents both computational and theoretical challenges. It is common to use conditional autoregressive (CAR) or Markov random field (MRF) priors to characterize spatial dependencies among spatially connected voxels (Besag, 1986; Banerjee, Carlin, and Gelfand, 2004), but estimating spatial correlation for a large number voxels, which ranges from ten thousands to more than 500,000, in the 3D volume (or 2D surface) is computationally prohibited. Moreover, it can be restrictive to assume a specific type of correlation structure, such as CAR and MRF, for the whole 3D volume (or 2D surface). Although the region-of-interest (ROI) method based on anatomically defined ROIs can model the spatial correlation among these ROIs, it essentially ignores the spatial correlation structure in the neighboring voxels within each ROI (Bowman, 2007). Moreover, the ROI method is also based on a stringent assumption that all voxels in the same ROI are homogeneous, which is largely false.

1.3 Multiscale Adaptive Generalized Estimating Equations for Longitudinal Neuroimaging Data

The primary goal of a longitudinal neuroimaging study is to characterize individual change in neuroimaging measurements (e.g., volumetric and morphometric) over time,

and the covariates of interest, such as age, diagnostic status, and gender, that influence change (Whitwell, 2008). A distinctive feature of longitudinal neuroimaging data is that neuroimaging data have a temporal order. Imaging measurements of the same individual usually exhibit positive correlation and the strength of the correlation decreases with the time separation. Ignoring temporal correlation structure in imaging measures likely would influence subsequent statistical inference, such as increasing false positive and negative errors, and lead to misleading scientific inference (Diggle, Heagerty, Liang and Zeger 2002; Fitzmaurice, Laird, and Ware 2004).

Many large-scale longitudinal imaging studies including the Alzeimer's Disease Neuroimaging Initiaitve (ADNI) and the NIH MRI study of normal brain have been or are being widely conducted to better understand the progress of neuropsychiatric and neurodegenerative diseases or the normal brain development/aging (Evans, and B.D.C. Group, 2006; Almli, Rivkin, and McKinstry, 2007; Hua et al., 2009; Fan et al., 2008). ADNI as one major ongoing neuroimaging longitudinal study is to search for the neuroimaging biomarkers for cognitive changes associated with Mild Cognitive Impairment and Alzheimer's Disease (Hua et al., 2009; Fan et al., 2008). However, analysis of these longitudinal imaging data has been hindered by the lack of advanced image processing and statistical tools for analyzing complex and correlated imaging data along with behavioral and clinical data. Recently, cross-sectional image processing and voxel-wise methods have been developed and used, but they are in general not optimal in power. For instance, the popular neuroimaging software platforms including AFNI, statistical parametric mapping (SPM) and FMRIB Software Library (FSL) cannot serve the emerging needs of these projects for voxel based longitudinal analysis.

Chapter 2

Regression Models for Identifying Noise Sources in Magnetic Resonance Images

2.1 Introduction

MRI is an non-invasive imaging technique used extensively for clinical diagnosis and medical research. MRIs, however, contain varying amounts of noise of diverse origins, including noise from stochastic variation, numerous physiological processes, eddy currents, artifacts from the differing magnetic field susceptibilities of neighboring tissues, rigid body motion, non-rigid motion, and many others (Huettel, Song, and McCarthy 2004). Some noise components, including bulk motion from cardiac pulsation and head or body movement, generate unusual observations, or statistical 'outliers', that differ substantially from most MR data that do not contain those noise sources (at least, not to the same degree). Previous studies have shown that those noise components can introduce substantial bias into measurements and estimation made from those images, such as indices for the principle direction of fiber tracts in diffusion tensor images (Skare, Li, Nordell, and Ingvar 2000; Luo and Nichols 2003; Nowark 1999). Identifying and reducing

these noise components in MR images is essential to improving the validity and accuracy of studies designed to map the structure and function of the human body.

The Rician distribution will be shown below as the model for characterizing the stochastic noise in the magnitude of MR data. Formal assessment of the quality of MR images should include identification of non-stochastic noise components as well, such as those from susceptibility artifacts and rigid body motion. These non-stochastic noise sources usually introduce statistical outliers in some or all of the volume elements, called "voxels", of the image, the elemental units from which an image is constructed. Diagnostic procedures, such as an analysis of residuals, can be useful tools for detecting discrepancies between those outliers and other observations at all voxels. Moreover, even under the sole presence of stochastic noise, diagnostic methods are valuable for detecting discrepancies between MR data and fitted models at the voxel level. Such discrepancies can be caused by partial volume effects in the MR image (i.e., the presence of multiple tissues in the same volume element, or voxel in the tissue that corresponds with the given pixel in the image). In diffusion tensor images (DTIs), for instance, modeling these effects in voxels having multiple tissue compartments can be vitally important for reconstructing complex tissue structure in the human brain in vivo (Tuch, Reese, Wiegell, Makris, Belliveau, Wedeen 2002; Alexander, Barker, and Arridge 2002).

The aim of this paper is to introduce a Rician regression model and its related normal models to characterize noise contributions in various MRI modalities and to develop its associated estimation methods and diagnostic tools. We develop the estimation algorithms for calculating the maximum likelihood estimates of three regression models for MRI data. We develop a diagnostic procedure to systematically assess the quality of MR images using a variety of diagnostic techniques, including an analysis of residuals, Cook's distance, goodness-of-fit test statistics, influence measures, and graphical analyses. We use the p-values of test statistics to evaluate directly the goodness of fit of the fitted regression models to the MRI data. Two diagnostic measures, standardized

residuals and Cook's distance, identify in each voxel of the image outliers that can be caused by motion artifacts and other noise components. Graphical tools include three-dimensional (3D) images of statistical measures that can isolate problematic voxels, as well as two-dimensional (2D) plots for assessing the compatibility of the fitted regression model with data in individual voxels. Finally, we apply these diagnostic techniques to diffusion tensor images and demonstrate that the techniques are able to identify subtle artifacts and experimental variation not captured by the Rician model.

We will next present the Rician regression model and its two related normal models and discuss some of their statistical properties. Estimation algorithms will be used to maximize the likelihood function of the regression models proposed. Then we will develop diagnostic procedures consisting of goodness-of-fit statistics, influence measures, and graphical analyses. Simulation studies will assess the empirical performance of the estimation algorithms and goodness-of-fit statistics under different experimental conditions. Finally, we will analyze a real data set to illustrate an application of these methods, before offering some concluding remarks.

2.2 The Regression Models for MR Images

2.2.1 Model Formulation

We usually acquire n MR images for each subject. Each MRI contains N voxels, and thus each voxel contains n measurements. We use $\{(S_i, x_i) : i = 1, \cdots, n\}$ to denote the n measurements at a single voxel, where S_i denotes the MRI signal intensity and x_i includes all the covariates of interest, such as the gradient directions and gradient strengths for acquiring diffusion tensor images. In MR images, $S_i = \sqrt{R_i^2 + I_i^2}$ and ϕ_i are, respectively, the magnitude and phase of a complex number (R_i, I_i) from data in the imaging domain such that $R_i = S_i \sin(\phi_i)$ and $I_i = S_i \cos(\phi_i)$ for $i = 1, \cdots, n$.

The MR signal S_i is assumed to follow a Rician distribution with parameters μ_i and

σ^2, denoted by $S_i \sim \mathrm{R}(\mu_i, \sigma^2)$, under the presence solely of stochastic noise (Rice 1945). Suppose that R_i and I_i are independent and follow normal distributions with the same variance σ^2, and with means $\mu_{R,i}$ and $\mu_{I,i}$, respectively. Thus, the joint density function of (S_i, ϕ_i) can be written as

$$p(S_i, \phi_i) = \frac{S_i}{2\pi\sigma^2} \exp\{-0.5\sigma^{-2}(S_i \sin(\phi_i) - \mu_{R,i})^2 - 0.5\sigma^{-2}(S_i \cos(\phi_i) - \mu_{I,i})^2\}.$$

Integrating out ϕ_i, we obtain the density function of the Rician distribution as follows:

$$p(S_i|\mu_i, \sigma^2) = \frac{S_i}{\sigma^2} \exp\{-0.5\sigma^{-2}(S_i^2 + \mu_i^2)\} I_0\left(\frac{\mu_i S_i}{\sigma^2}\right) 1(S_i \geq 0), \tag{2.1}$$

where $\mu_i = \sqrt{\mu_{R,i}^2 + \mu_{I,i}^2}$, $1(\cdot)$ is an indicator function, and $I_0(z) = \int_0^{2\pi} \exp(z \cos\phi) d\phi/(2\pi)$ denotes the 0th order modified Bessel function of the first kind (Abramowitz and Stegun 1965).

We formally define a *Rician regression* model by assuming that

$$S_i \sim \mathrm{R}(\mu_i(\beta), \sigma^2) \quad \text{and} \quad \mu_i(\beta) = f(x_i, \beta), \tag{2.2}$$

where β is a $p \times 1$ vector in R^p and $f(\cdot, \cdot)$ is a known link function, which depends on the particular MR imaging modalities (e.g., anatomical, functional, DTI, etc). Because the density in (2.1) does not belong to the exponential family, the Rician regression model is not a special case of a generalized linear model (McCullagh and Nelder 1989).

We calculate the kth moment of S_i given x_i as follows. Let $I_k(z)$ be the k-th modified Bessel function of the first kind (Abramowitz and Stegun 1965) defined by $I_k(z) = \int_0^{2\pi} \cos(k\phi) e^{z\cos\phi} d\phi/(2\pi)$. It can be shown that the kth moment of S_i given x_i (Sijbers, den Dekker, Scheunders, and van Dyck 1998a) is calculated as

$$\mathrm{E}(S_i^k|x_i) = (2\sigma^2)^{k/2}\Gamma(1 + \frac{k}{2}) M\left(-\frac{k}{2}; 1; -\frac{\mu_i(\beta)^2}{2\sigma^2}\right), \tag{2.3}$$

where $\Gamma(\cdot)$ is the Gamma function and $M(\cdot)$ is the Kummer function (or confluent hypergeometric function) (Abramowitz and Stegun, 1965). The even moments of S_i given x_i are simple polynomials. For instance,

$$\mathrm{E}(S_i^2|x_i) = \mu_i(\beta)^2 + 2\sigma^2 \quad \text{and} \quad \mathrm{E}(S_i^4|x_i) = \mu_i(\beta)^4 + 8\sigma^2\mu_i(\beta)^2 + 8\sigma^4. \tag{2.4}$$

However, the odd moments of S_i given x_i are much more complex; for instance,

$$\mathrm{E}(S_i|x_i) = \sigma\sqrt{\frac{\pi}{2}}\exp\{-\frac{\mu_i(\beta)^2}{4\sigma^2}\}\left[\left(1+\frac{\mu_i(\beta)^2}{2\sigma^2}\right)I_0\left(\frac{\mu_i(\beta)^2}{4\sigma^2}\right)+\frac{\mu_i(\beta)^2}{2\sigma^2}I_1\left(\frac{\mu_i(\beta)^2}{4\sigma^2}\right)\right]. \tag{2.5}$$

The Rician distribution can be well approximated by a normal distribution at high signal-to-noise ratios (SNR), defined by $\mu_i(\beta)/\sigma$. When SNR$\leq$ 1, the Rician distribution is far from being Gaussian. When SNR$\geq$ 2, $R(\mu_i(\beta), \sigma^2)$ can be closely approximated by a *normal regression* model (Gudbjartsson and Patz 1995) (Fig. 2.1a), which is given by

$$S_i \sim N(\sqrt{\mu_i(\beta)^2 + \sigma^2}, \sigma^2) \;\text{ and }\; \mu_i(\beta) = f(\mathbf{x}_i, \beta). \tag{2.6}$$

Moreover, the second moment of $R(\mu_i(\beta), \sigma^2)$ equals that of $N(\sqrt{\mu_i(\beta)^2 + \sigma^2}, \sigma^2)$, while $\mathrm{E}(S_i|x_i)$ in (2.5) can be accurately approximated by $\sqrt{\mu_i(\beta)^2 + \sigma^2}$ even when SNR is close to 1 (Fig. 2.1b). Furthermore, if SNR is greater than 5, then $\sqrt{\mu_i(\beta)^2 + \sigma^2} = \mu_i(\beta)\sqrt{1 + 1/\mathrm{SNR}^2} \approx \mu_i(\beta)$. Thus, $R(\mu_i(\beta), \sigma^2)$ can be approximated by another normal regression model given by

$$S_i \sim N(\mu_i(\beta), \sigma^2) \;\text{ and }\; \mu_i(\beta) = f(\mathbf{x}_i, \beta). \tag{2.7}$$

2.2.2 Examples

The regression models proposed here include statistical models for various MRI modalities, including DTI and functional MRI. For the purposes of illustration, we consider

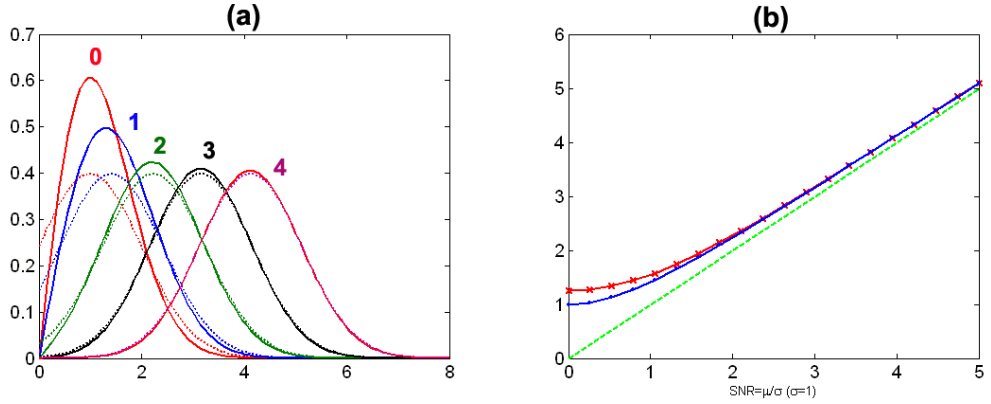

Figure 2.1: Rician distribution: (a) $R(\mu, 1)$ and $N(\sqrt{\mu^2+1}, 1)$ for $\mu = 0, 1, 2, 3, 4$; (b) the mean functions of $R(\mu, 1)$ (red), $N(\sqrt{\mu^2+1}, 1)$ (blue) and $N(\mu, 1)$ (green) for $\mu \in [0, 5]$.

the following five examples.

EXAMPLE 1. Stochastic noise in MRI data follows a $R(0, \sigma^2)$ distribution, which is a highly skewed Rayleigh distribution. The first two moments of $R(0, \sigma^2)$ are given by $\mathrm{E}(S_i|x_i) = \sigma\sqrt{0.5\pi}$ and $\mathrm{E}(S_i^2|x_i) = 2\sigma^2$. Without any other noise components present, such as ghosting artifacts, we can use the MR data in the background of the image to estimate σ^2. However, under the presence of non-stochastic noise components, such as ghosting artifacts, the background MR signals do not follow a Rician distribution, and the estimate of σ^2 is usually a biased estimate of σ^2. Therefore, testing whether the MR signal in a single voxel truly follows a Rician model is useful to detect the presence of non-stochastic noise components.

EXAMPLE 2. If we apply an inversion snapshot FLASH imaging sequence to measure T_1 relaxation times, then we have $\mu_i(\beta) = \rho\left(1 - 2\exp\left(-t_i T_1^{-1}\right)\right)$, where x_i is time t_i and β includes a pseudo proton density ρ and spin-lattice or longitudinal relaxation constant T_1. It has been shown that the use of the Rician model leads to a substantial increase in precision of the estimated T_1 (Karlsen, Verhagen, and Bovee 1999).

If the decay of transverse magnetization is mono-exponential and conventional spin-echo imaging is used, then $f(x_i, \beta)$ is given by $\mu_i(\beta) = \rho\exp\left(-\mathrm{TE}_i \times {T_2}^{-1}\right)$, where x_i is the echo time TE_i and $\beta = (\rho, T_2)$, in which T_2 is the spin-spin relaxation constant.

EXAMPLE 3. In a fMRI session, fMRI volumes are acquired repeatedly over time while a subject performs a cognitive or behavioral task. Over the course of the experiment, n fMRI volumes are typically recorded at acquisition times $t_1, \cdots, t_n$. The standard method for computing the statistical significance of task-related activations is to use only the magnitude MR image at time t_i for $i = 1, \cdots, n$. The magnitude image at time t_i follows a Rician distribution with $\mu_i(\beta) = x_i^T \beta$, the superscript T denotes transpose and x_i may include responses to differing stimulus types, the rest status, and various reference functions (Rowe and Logan 2005; den Dekker and Sijbers 2005).

EXAMPLE 4. Diffusion tensor images (DTI) have been widely used to reconstruct the pathways of white matter fibers in the human brain in vivo (Basser, Mattiello, and LeBihan 1994 a, b; Xu et al. 2002). A single shot echo-planar imaging (EPI) technique is often used to acquire DWIs with moderate resolution (e.g., 2.5 mm×2.5 mm× 2.5 mm), and then diffusion tensors can estimated using DWI data. In voxels with a single fiber population, a simple diffusion model assumes that

$$\mu_i(\beta) = S_0 \exp(-b_i r_i^T D r_i) \tag{2.8}$$

for $i = 1, \cdots, n$, where $x_i = (b_i, r_i, t_i)$, in which t_i is the acquisition time for the ith image, $r_i = (r_{i,1}, r_{i,2}, r_{i,3})^T$ is an applied gradient direction and b_i is the corresponding gradient strength. In addition, S_0 is the signal intensity in the absence of any diffusion-weighted gradient and the diffusion tensor $D = (D_{i,j})$ is a 3×3 positive definite matrix. The three eigenvectors of D constitute the three diffusion directions and the corresponding eigenvalues define the degrees of diffusivity along each of the three spatial directions. Many tractography algorithms attempt to reconstruct fiber tracts by connecting spatially consecutive eigenvectors corresponding to the largest eigenvalues of the diffusion tensors (DTs) across adjacent voxels.

The SNRs in DW images are relatively low. The DW imaging acquisition scheme usually consists of few baseline images with $b = 0\text{s/mm}^2$ and many DW images with

$b-$values greater than zero. As an illustration, we selected a representative subject from an existing DTI data set and calculated the estimates of S_0/σ and eigenvalues of D, denoted by $\lambda_1 \geq \lambda_2 \geq \lambda_3$, in all voxels containing anisotropic tensors (λ_1 was much larger than λ_3) (Fig. 2.2a and 2.2b). For these anisotropic tensors, SNR$= S_0/\sigma$ in baseline images varied from 0 to 15 with a mean close to 6 (Fig. 2.2c), while λ_1 varied from 0.5 (10^{-3} mm^2/s) to 2.0 (10^{-3}mm^2/s) with a mean close to 1.0 (10^{-3} mm^2/s). For a moderate gradient strength $b_i \approx 1000$s/mm^2, SNR$= \exp(-b_i r_i^T D r_i) \times (S_0/\sigma)$ in DWIs varied from 0 to 8 with a mean close to 2.5 (Fig. 2.2d).

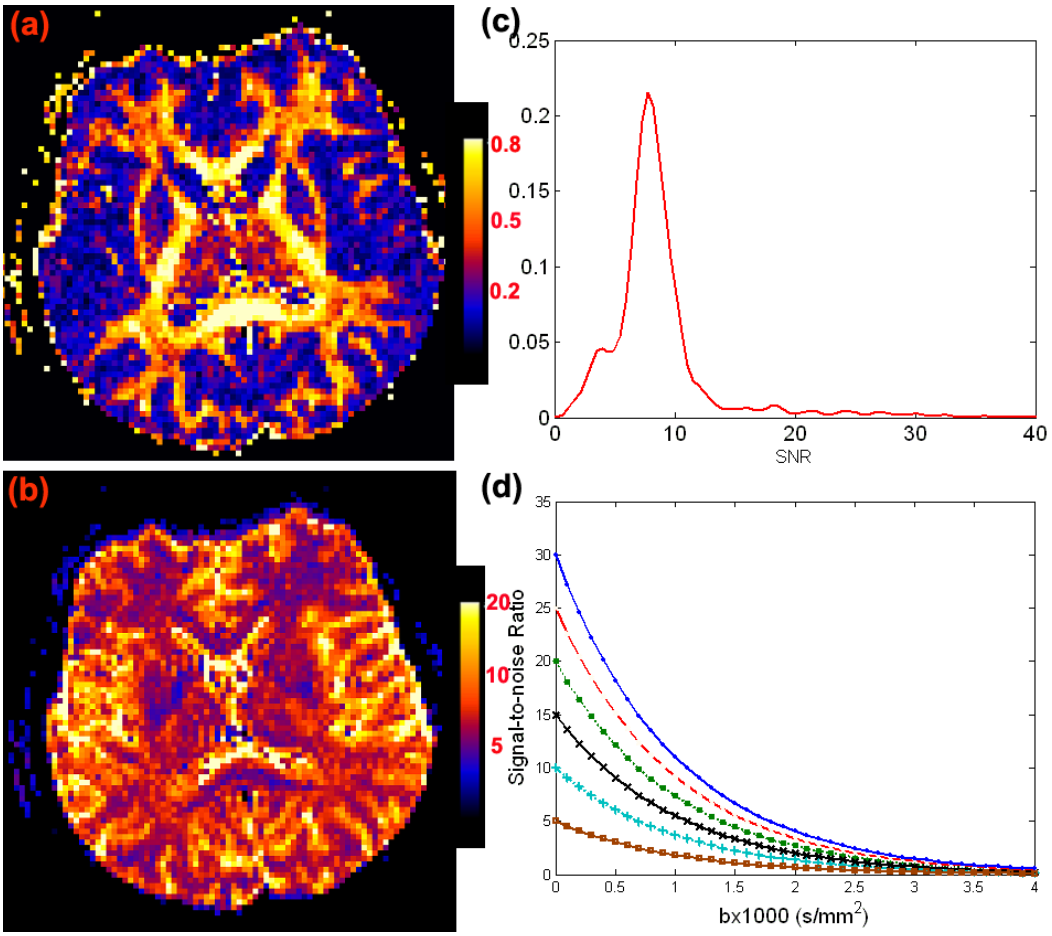

Figure 2.2: Maps of (a) FA; (b) S_0/σ; (c) the kernel density of S_0/σ values for anisotropic tensors having FA$\geq$ 0.5 at a selective slice from a single subject; and (d) the signal-to-noise ratio $S_0 \exp(-b_i)/\sigma$ as a function of b_i ($\times$1000 s/mm^2) at each $S_0/\sigma \in \{5, 10, 15, 20, 25, 30\}$.

To account for the presence of multiple fibers within a single voxel, a diffusion model

with M compartments may be written as

$$\mu_i(\beta) = S_0 \sum_{k=1}^{M} p_k \exp(-b_i r_i^T D_k r_i), \tag{2.9}$$

where p_k denotes the proportion of each compartment such that $\sum_{k=1}^{M} p_k = 1$ and $p_k \geq 0$ and where D_k is the diffusion tensor for the kth compartment. Recent studies have shown that elucidating multiple fibers need large b values (Tuch et al. 2002; Alexander, Barker, and Arridge 2002; Jones and Basser 2004). For instance, Alexander and Barker (2005) have shown that the optimal values of b for recovering two fibers are in the range [2200, 2800]s/mm^2. For large b values, SNR in DWIs can be very close to zero (Fig. 2.2d).

Example 5. If we are only interested in the apparent diffusion coefficient (ADC) normal to the fiber direction in white matter, then we can use a single EPI technique to acquire MR images based on multiple b_i factors in the absence of a diffusion-weighted gradient (Kristoffersen 2007). A simple mono-exponential diffusion model assumes that $\mu_i(\beta) = S_0 \exp(-b_i d)$ for $i = 1, \cdots, n$. The values of ADC are in the range of $[0.2, 3]$ ($\times 10^{-3}$mm^2/s) for the human brain. Furthermore, a diffusion model with M compartments may be written as $\mu_i(\beta) = S_0 \sum_{k=1}^{M} p_k \exp(-b_i d_k)$.

2.2.3 Estimation methods

We consider estimation algorithms for the two normal models (2.6) and (2.7). Because the normal model (2.7) is a standard nonlinear regression model, we can directly use the standard Levenberg-Marquardt method to calculate the maximum likelihood estimate of θ. For the normal model (2.6), we propose an iterative procedure to maximize its log-likelihood function given by

$$\ell(\beta, \sigma^2) = -0.5n \log \sigma^2 - 0.5 \sum_{i=1}^{n} \{S_i - \sqrt{\mu_i(\beta)^2 + \sigma^2}\}^2 / (\sigma^2).$$

We use the Levenberg-Marquardt method to minimize $\sum_{i=1}^n \{S_i - \mu_i(\beta)\}^2$, which yields an initial estimator $\beta^{(0)}$, and we subsequently calculate $(\sigma^2)^{(0)} = \sum_{i=1}^n \{S_i - \mu_i(\beta^{(0)})\}^2/n$. Given $(\sigma^2)^{(r)}$, we use the Levenberg-Marquardt method to calculate $\beta^{(r+1)}$ that minimizes $\sum_{i=1}^n \{S_i - \sqrt{\mu_i(\beta)^2 + (\sigma^2)^{(r)}}\}^2$. Conditional on $\beta^{(r+1)}$, we use the Newton-Raphson algorithm to calculate $\sigma^{(r+1)}$ by maximizing $\ell(\beta^{(r+1)}, \sigma^2)$. This iterative algorithm stops when the absolute difference between consecutive $\theta^{(t)}$s is smaller than a predefined small number, say 10^{-4}.

We introduce an efficient EM algorithm (Dempster, Laird and Rubin 1977) for maximizing the likelihood function of the Rician model (2.2). The key idea is to introduce a latent phase variable $\phi_i \in [0, 2\pi]$ for each S_i such that the joint density of (S_i, ϕ_i) is given by

$$p(S_i, \phi_i | x_i) = \frac{1}{2\pi\sigma^2} S_i \exp\left(-\frac{\mu_i(\beta)^2 + S_i^2 - 2S_i\mu_i(\beta)\cos(\phi_i)}{2\sigma^2}\right).$$

Let $Y_o = (S_1, x_1, \cdots, S_n, x_n)$ denote the observed data and $Y_m = (\phi_1, \cdots, \phi_n)$ denotes the missing data. The log-likelihood function of $Y_c = (Y_o, Y_m)$, defined by $L_c(\theta|Y_c)$, can be written as

$$-n\log(2\pi\sigma^2) + \sum_{i=1}^n \log S_i - 0.5\sigma^{-2}\sum_{i=1}^n \{\mu_i^2(\beta) + S_i^2 - 2S_i\mu_i(\beta)\cos(\phi_i)\}. \tag{2.10}$$

A standard EM algorithm consists of two steps: the expectation (E) step and the maximization (M) step as follows. The E-step evaluates $Q(\theta|\theta^{(r)}) = \mathrm{E}\{L_c(\theta|Y_c)|Y_o, \theta^{(r)}\}$, where the expectation is taken with respect to the conditional distribution $p(Y_m|Y_o, \theta^{(r)}) = \prod_{i=1}^n p(\phi_i|S_i, \theta^{(r)})$. We can show that

$$p(\phi_i|S_i, \theta) = \frac{1}{2\pi I_0(\sigma^{-2}S_i\mu_i(\beta))} \exp\{\sigma^{-2}S_i\mu_i(\beta)\cos(\phi_i)\} 1(\phi_i \in [0, 2\pi]).$$

Thus, $Q(\theta|\theta^{(r)})$ is given by

$$-n\log(\sigma^2) - 0.5\sigma^{-2}\sum_{i=1}^{n}\left\{\mu_i^2(\beta) + S_i^2 - 2S_i\mu_i(\beta)W_i(\theta^{(r)})\right\}, \tag{2.11}$$

where $W_i(\theta) = I_1(\sigma^{-2}f(x_i,\beta)S_i)/I_0(\sigma^{-2}f(x_i,\beta)S_i)$.

The M-step is to determine the $\theta^{(r+1)}$ that maximizes $Q(\theta|\theta^{(r)})$. However, because the M-step does not have a closed form, $\theta^{(r+1)}$ is obtained via two conditional maximization steps (Meng and Rubin 1993). Given $\beta^{(r)}$, we can derive

$$(\sigma^2)^{(r+1)} = 0.5n^{-1}\sum_{i=1}^{n}\left\{\mu_i^2(\beta^{(r)}) + S_i^2 - 2S_i\mu_i(\beta^{(r)})W_i(\theta^{(r)})\right\}.$$

Conditional on $(\sigma^2)^{(r+1)}$, we can determine $\beta^{(r+1)}$ by minimizing $G(\beta|\beta^{(r)}) = \sum_{i=1}^{n}\{\mu_i(\beta) - W_i(\theta^{(r)})S_i\}^2$. This is a standard nonlinear least squares problem, to which the Levenberg-Marquardt method can be applied. Furthermore, we may employ a generalized EM algorithm, in which the E-step is unchanged, but we replace the M-step with a generalized M-step to identify a $\beta^{(r+1)}$ such that $G(\beta^{(r+1)}|\beta^{(r)}) \leq G(\beta^{(r)}|\beta^{(r)})$. Under mild conditions, the sequence $\{\theta^{(r)}\}$ obtained from the EM algorithm converges to the maximum likelihood estimate, denoted by $\hat{\theta}$ (Meng and Rubin 1993).

The next important issue is to evaluate the covariance matrix of $\hat{\theta}$, which can be obtained by inverting either the Hessian matrix or the Fisher information matrix of the observed-data log-likelihood function. For instance, for the normal model (2.6), it is straightforward to calculate the second derivative of $\ell(\beta,\sigma^2)$. For the Rician model (2.2), we use the missing information principle (Louis 1982). Calculation of the first and second derivatives of $L_c(\theta|Y_c)$ with respect to θ is straightforward and hence is omitted here for brevity.

2.3 A Diagnostic Procedure

We propose a diagnostic procedure to identify noise components in MR images at all levels of the SNR. Our diagnostic procedure has three major components: (a) the use of goodness-of-fit test statistics to test the assumptions of the Rician model across all voxels of the image; (b) the use of influence measures to identify outliers; (c) the use of 2D and 3D graphs to search for various artifacts and to detect intravoxel variability. At a high SNR, these diagnostic measures of the Rician model reduce to those of the normal models (2.6) and (2.7). Thus, we will not specifically develop diagnostic measures of the two normal models. Furthermore, in the normal models (2.6) and (2.7), the goodness-of-fit statistics developed here are completely new.

2.3.1 Goodness-of-fit test statistics

We develop test statistics to check model misspecification in the Rician model (2.2). These test statistics are valuable for revealing two kinds of challenges in working with MR images. The first is to identify those voxels in which the MR signal contains substantial noise components that are other than stochastic noise. The second challenge is to identify those voxels in which the signal is affected strongly by partial volume effects.

Thus, we are interested in testing whether $f(x_i, \beta)$ is correctly specified. In most statistical models including generalized linear models, testing the specification of the link function is equivalent to testing the mean structure of the response variable (Stute 1997). However, because, in the Rician model (2.2), $E(S_i|x_i)$ does not have a simple form, testing directly the mean structure of the response is likely to be tedious and difficult. Let $W(\theta) = I_1(B(\theta))/I_0(B(\theta))$, where $B(\theta) = \sigma^{-2} f(x, \beta)S$. We also note the simple equality $E[W(\theta)S|x] = f(x, \beta)$, when the Rician model (2.2) is correctly specified. Thus, we suggest testing $h(\theta) = E[W(\theta)S|x] - f(x, \beta) = 0$, for which the null

and alternative hypotheses are stated as follows:

$$H_0^{(1)} : h(\theta) = 0 \text{ for some } \theta \in \Theta \quad \text{versus} \quad H_1^{(1)} : h(\theta) \neq 0 \text{ for all } \theta \in \Theta. \tag{2.12}$$

Because $W(\theta)$ is close to one at a high SNR, testing $H_0^{(1)}$ is essentially testing whether $E(S|x) = f(x, \beta)$ in the normal model (2.7).

To test $H_0^{(1)}$, we develop two test statistics as follows. The first of these, the conditional Kolmogorov test (CK), is

$$\mathrm{CK}_1 = \sup_u |T_1(u; \hat{\theta})|, \tag{2.13}$$

where $T_1(u; \hat{\theta})$ is defined as

$$T_1(u; \hat{\theta}) = n^{-1/2} \sum_{i=1}^{n} 1(x_i^T \hat{\beta} \leq u)[W_i(\hat{\theta}) S_i - \mu_i(x_i, \hat{\beta})]. \tag{2.14}$$

Under the null hypothesis, $E[T_1(u; \theta_*)]$ should be close to zero, where $\theta_* = (\beta_*, \sigma_*^2)$ is the true value of θ. Therefore, a large value of CK_1 leads to rejection of the null hypothesis $H_0^{(1)}$.

We must derive the asymptotic null distribution of CK_1 to test rigorously whether $H_0^{(1)}$ is true. We regard $T_1(u; \hat{\theta})$ as a stochastic process indexed by $u \in R$. We can show that under $H_0^{(1)}$, as $n \to \infty$,

$$T_1(u; \hat{\theta}) = T_1(u; \theta_*) + \partial_\theta T_1(u; \theta_*)(\hat{\theta} - \theta_*) + o_p(1) = T_1(u; \theta_*) + \Delta_1(u)\sqrt{n}(\hat{\theta} - \theta_*) + o_p(1),$$

where $\Delta_1(u)$ is defined by

$$\Delta_1(u) = \int [\partial_\theta W(\theta_*) S - \partial_\theta f(x, \beta_*)] \, 1(x^T \beta_* \leq u) p(S|x, \theta_*) p(x) dS dx.$$

Moreover, using the central limit theorem (van der Vaart and Wellner 1996), we can

show that

$$\sqrt{n}(\hat{\theta} - \theta_*) = n^{-1/2} \sum_{i=1}^{n} \psi(S_i, x_i; \theta_*) + o_p(1), \tag{2.15}$$

where $\psi(\cdot, \cdot; \theta_*)$ is a known influence function depending on the likelihood function of the Rician model (2.2). Finally, using empirical process theory (van der Vaart and Wellner 1996), we can show that the asymptotic null distribution of CK_1 depends on the asymptotic distribution of $(T_1(\cdot, \theta_*), \sqrt{n}(\hat{\theta} - \theta_*)^T)^T$, which is given in Theorem 1.

The second test statistic that we propose is based on

$$T_2(\alpha, u; \hat{\theta}) = n^{-1/2} \sum_{i=1}^{n} [W_i(\hat{\theta}) S_i - \mu_i(\hat{\beta})] 1(x_i^T \alpha \leq u), \tag{2.16}$$

where $\Pi = \{\alpha \in R^d : \alpha^T \alpha = 1\} \times [-\infty, \infty]$. Following the reasoning in Escanciano (2006), we can show that $H_0^{(1)}$ is equivalent to testing

$$E\{[W_i(\theta) S_i - \mu_i(\beta)] 1(x^T \alpha \leq u)\} = 0 \tag{2.17}$$

for almost every $(\alpha, u) \in \Pi$ for some $\theta_* \in \Theta$. Let $F_{n,\alpha}(u)$ be the empirical distribution function of $\{\alpha^T x_i : i = 1, \cdots, n\}$. Then, we define the Cramer-von Mises test statistic as follows:

$$\mathrm{CM}_1 = \int_{\Pi} T_2(\alpha, u; \hat{\theta})^2 F_{n,\alpha}(du) d\alpha, \tag{2.18}$$

where $d\alpha$ is taken with respect to the uniform density on the unit sphere. A simple algorithm for computing CM_1 can be found in Escanciano (2006). A large value of CM_1 leads to rejection of $H_0^{(1)}$. Similar to CK_1, we can show that $T_2(\alpha, u; \hat{\theta})$ is approximated as

$$T_2(\alpha, u; \hat{\theta}) = T_2(\alpha, u; \theta_*) + \Delta_2(\alpha, u) \sqrt{n}(\hat{\theta} - \theta_*) + o_p(1),$$

where $\Delta_2(\alpha, u) = \int [\partial_\theta W(\theta_*) S - \partial_\theta f(x, \beta_*)] \, 1(\alpha^T x \leq u) p(S|x, \theta_*) p(x) dS dx$. Therefore, the asymptotic null distribution of CM_1 depends on the asymptotic distribution of $(T_2(\alpha, u; \theta_*), \sqrt{n}(\hat{\theta} - \theta_*)^T)^T$, which is also given in Theorem 1. The detailed proof of

Theorem 1 can be found in a supplementary report. We are now led to the following theorem.

THEOREM 1. Under the null hypothesis $H_0^{(1)}$, we have the following results:

i) $\sqrt{n}(\hat{\theta} - \theta_*) = n^{-1/2} \sum_{i=1}^{n} \psi_{n,i} + o_p(1)$.

ii) $(T_1(\cdot;\theta_*), \sqrt{n}(\hat{\theta}-\theta_*)^T)^T$ converges in distribution to $(G_1(\cdot;\theta_*), \nu_1^T)^T$, where $(G_1(\cdot;\theta_*), \nu_1^T)$ is a Gaussian process with mean zero and covariance function $C_1(u_1, u_2)$, which is given by

$$C_1(u_1,u_2) = \int\int \begin{pmatrix} [W(\theta_*)S_i - f(x,\beta_*)]1(x^T\beta_* \le u_1) \\ \psi(S,x;\theta_*) \end{pmatrix} \times \begin{pmatrix} [W(\theta_*)S - f(x,\beta_*)]1(x^T\beta_* \le u_2) \\ \psi(S,x;\theta_*) \end{pmatrix}^T p(S|x,\theta_*)dSdp(x). \tag{2.19}$$

iii) CK_1 converges in distribution to $\sup_u |T_1(u;\theta_*) + \Delta_1(u)^T\nu_1|$.

iv) $(T_2(\cdot,\cdot;\theta_*), \sqrt{n}(\hat{\theta} - \theta_*)^T)^T$ converges in distribution to $(G_2(\cdot,\cdot;\theta_*), \nu_1^T)^T$, where $(G_2(\cdot,\cdot;\theta_*), \nu_1^T)$ is a Gaussian process with mean zero and covariance function $C_2((\alpha_1,u_1), (\alpha_2,u_2))$, which is given by

$$C_2((\alpha_1,u_1),(\alpha_2,u_2)) = \int\int \begin{pmatrix} [W(\theta_*)S - f(x,\beta_*)]1(x^T\alpha_1 \le u_1) \\ \psi(S,x;\theta_*) \end{pmatrix} \times \begin{pmatrix} [W(\theta_*)S - f(x,\beta_*)]1(x^T\alpha_2 \le u_2) \\ \psi(S,x;\theta_*) \end{pmatrix}^T p(S|x,\theta_*)dSdp(x). \tag{2.20}$$

v) CM_1 converges in distribution to $\int_\Pi |T_2(\alpha,u;\theta_*) + \Delta_2(\alpha,u)\nu_1|^2 F_\alpha(du)d\alpha$, where $F_\alpha(u)$ is the true cumulative distribution function of $\alpha^T x$.

Theorem 1 characterizes the limiting distributions of CK_1 and CM_1 under the null hypotheses.

Because $E(S_i^2|x_i)$ has a simple form, we further use the second moment of S_i given

x_i to test the specification of the link function. Specifically, the null and alternative hypotheses are given by

$$H_0^{(2)} : E(S^2|x) = f(x,\beta)^2 + 2\sigma^2 \text{ for some } \theta \in \Theta,$$
$$H_1^{(2)} : E(S^2|x) \neq f(x,\beta)^2 + 2\sigma^2 \text{ for all } \theta \in \Theta.$$

Similar to testing $H_0^{(1)}$ against $H_1^{(1)}$, we introduce two other stochastic processes given by

$$\begin{aligned} T_3(u;\theta) &= n^{-1/2}\sum_{i=1}^{n} 1(x_i^T\beta \leq u)[S_i^2 - \mu_i(\beta)^2 - 2\sigma^2] \text{ and} \\ T_4(\alpha,u;\theta) &= n^{-1/2}\sum_{i=1}^{n}[S_i^2 - \mu_i(\beta)^2 - 2\sigma^2]1(x_i^T\alpha \leq u). \end{aligned}$$

Based on $T_3(u;\theta)$ and $T_4(\alpha,u;\theta)$, we can develop two additional test statistics:

$$\text{CK}_2 = \sup_u |T_3(u;\hat{\theta})| \text{ and } \text{CM}_2 = \int_\Pi T_4(\alpha,u;\hat{\theta})^2 F_{n,\alpha}(du)d\alpha. \tag{2.21}$$

Similar to the reasoning in Theorem 1, we can establish the asymptotic null distributions of CK_2 and CM_2, which we therefore omit here. Because the normal model (2.6) has the same second moment as the Rician model (2.2), the test statistics CK_2 and CM_2 are valid for model (2.6) at all levels of the SNR. So far, we have introduced four test statistics CK_1, CK_2, CM_1, and CM_2, each of which may have different sensitivities in detecting the misspecification of a Rician model in various circumstances, which we will investigate with the simulation studies of Section 2.4.

We note two types of correlation existing in CK_1, CK_2, CM_1, and CM_2 at the local and global levels. At the local level, there may be strong correlations among these four test statistics in each voxel, because the same MRI data within the voxel are used to calculate them. At the global level, we calculate these four test statistics across multiple brain regions or across the many voxels of the imaging volume. MRI data in small spatial

neighborhoods show strong similarlity, whereas MRI data in voxels more distant from one another show less similarity. Thus, the same test statistics $\mathrm{CK}_1(d)$ (or $\mathrm{CK}_2(d)$, $\mathrm{CM}_1(d)$, and $\mathrm{CM}_2(d)$) are likely to be positively correlated in small spatial neighborhoods, where d denotes a particular voxel in an MRI. Finally, we need to compute the uncorrected and corrected p-values of these four test statistics at the local and global levels.

2.3.2 Resampling method

Although the asymptotic distributions of $\mathrm{CK}_1(d)$, $\mathrm{CK}_2(d)$, $\mathrm{CM}_1(d)$, and $\mathrm{CM}_2(d)$ have been derived in Theorem 1, these limiting distributions usually have complicated analytic forms. To alleviate this difficulty, we develop a resampling method to estimate the null distribution of the statistic $\mathrm{CK}_1(d)$ in each of the voxels in the MRI data. The next issue is to solve the issue of multiple testing. Because it is difficult to compute an accurate p-value of $\mathrm{CK}_1(d)$ at each voxel, we avoid use of the false discovery rate and choose to control the family-wise error rate based on the maxima of the $\mathrm{CK}_1(d)$ statistics defined by $\mathrm{CK}_{1,\mathcal{D}} = \max_{d\in\mathcal{D}}\mathrm{CK}_1(d)$, where $\mathcal{D}$ denotes the brain region. Specifically, we can easily extend the proposed resampling method to approximate the null distribution of the statistic $\mathrm{CK}_{1,\mathcal{D}}$. In the following, we will introduce voxel d into all of the notation, if necessary. Because we can develop similar methods for CK_2, CM_1, and CM_2, we avoid such repetition and simply present the six key steps in generating the stochastic processes that have the same asymptotic distribution as $\mathrm{CK}_1(d)$ and $\mathrm{CK}_{1,\mathcal{D}}$.

Step 1. Generate independent and identically distributed random variables, $\{v_i^{(q)} : i = 1, \cdots, n\}$, from a $N(0,1)$ distribution for $q = 1, \cdots, Q$, where Q is the number of replications, say $Q = 1000$.

Step 2. Calculate

$$T_1(u;\hat{\boldsymbol{\theta}}(d))^{(q)} = n^{-1/2}\sum_{i=1}^{n} v_i^{(q)}\{\mathrm{E}_i(\hat{\boldsymbol{\theta}}(d))\mathbf{1}(\mathbf{x}_i'\hat{\beta}(d) \le u) - \hat{\Delta}_1(d,u)\psi_{ni}(d)\}$$

where $\mathrm{E}_i(\hat{\theta}(d)) = W_i(\hat{\theta}(d))S_i - \mu(x_i, \hat{\beta}(d))$ and $\hat{\Delta}_1(d,u) = n^{-1}\sum_{i=1}^{n} \partial_\theta \mathrm{E}_i(\hat{\theta}(d))\mathbf{1}(\mathbf{x}_i'\hat{\beta}(d) \leq u)$. Note that conditional on the observed data, $T_1(u;\hat{\boldsymbol{\theta}}(d))^{(q)}$ converges weakly to the desired Gaussian process in Theorem 1 as $n \to \infty$ (van der Vaart & Wellner, 1996).

Step 3. Calculate the test statistics $\mathrm{CK}_1^{(q)}(d) = \sup_u |T_1(u;\hat{\theta}(d))^{(q)}|$ and $\mathrm{CK}_{1,\mathcal{D}}^{(q)} = \sup_{d\in\mathcal{D}} \mathrm{CK}_1^{(q)}(d)$ and obtain $\{\mathrm{CK}_1^{(q)}(d) : q = 1, \cdots, Q\}$ and $\{\mathrm{CK}_{1,\mathcal{D}}^{(q)} : q = 1, \cdots, Q\}$.

Step 4. Calculate the $p-$value of $\mathrm{CK}_1(d)$ using $\{\mathrm{CK}_1^{(q)}(d) : q = 1, \cdots, Q\}$.

Step 5. Calculate the $p-$value of $\mathrm{CK}_1(d)$ at each voxel d of the region according to $p(d) \approx Q^{-1}\sum_{q=1}^{Q} 1(\mathrm{CK}_1^{(q)}(d) \geq \mathrm{CK}_1(d))$.

Step 6. Calculate the corrected $p-$value of $\mathrm{CK}_1(d)$ at each voxel d of the region using $p_{\mathcal{D}}(d) \approx Q^{-1}\sum_{q=1}^{Q} 1(\mathrm{CK}_{1,\mathcal{D}}^{(q)} \geq \mathrm{CK}_1(d))$.

Finally, we present a plot of the uncorrected and corrected $-\log_{10}(p)$ values for our various test statistics, such as CM_1. Since the above procedure only requires the computation of all components of $T_1(u;\hat{\boldsymbol{\theta}}(d))$ once and the repeated calculation of $\mathrm{CK}_1^{(q)}(d)$, it is computationally efficient. To identify the precise source of noise that is responsible for misspecification of the model, we need to develop influence measures to quantify the influence of each data point at each voxel.

2.3.3 Influence measures

Next we develop two influence measures that identify in each voxel of an MR image statistical 'outliers' which exert undue influence on the estimation of the parameters and fitted values of the model. These influence measures are based on case-deletion diagnostics, which have been studied extensively in regression models (Cook and Weisberg 1982; Wei 1998). Influence measures for the Rician regression model, however, have not been developed previously. Therefore, we now discuss how to develop case-deletion measures for the Rician model.

Henceforth, we assume that σ^2 is a nuisance parameter and define $U(\beta) = (\mu_1(\beta), \cdots, \mu_n(\beta))^T$, $V(\theta) = \mathrm{diag}(V_1(\theta), \cdots, V_n(\theta))$, and $S_W(\theta) = (W_1(\theta)S_1, \cdots, W_n(\theta)S_n)^T$, where

$V_i(\theta) = \sigma^{-2}\mathrm{Var}(S_i W_i(\theta)) = -\sigma^{-2}\mu_i(\beta)^2 + E[\sigma^{-2}S_i^2 W_i(\theta)^2]$. Thus, the score function for β is given by $SC_n(\beta) = \sigma^{-2}D(\beta)^T V(\theta)e(\beta)$, where $D(\beta) = \partial U(\beta)/\partial\beta^T$ is an $n \times p$ matrix with the ith row $\partial\mu_i(\beta)/\partial\beta^T$ and $e(\theta) = V(\theta)^{-1}[S_W(\theta) - U(\beta)]$. Furthermore, the Fisher information matrix for β takes the form

$$F_n(\beta) = \sigma^{-2}\sum_{i=1}^{n}\frac{\partial\mu_i(\beta)}{\partial\beta}V_i(\theta)\frac{\partial\mu_i(\beta)}{\partial\beta^T} = \sigma^{-2}D(\beta)^T V(\theta)D(\beta).$$

To develop influence measures, we can write the maximum likelihood estimate of β as $\hat{\beta} = [D(\hat{\beta})^T V(\hat{\theta})D(\hat{\beta})]^{-1}D(\hat{\beta})^T V(\hat{\theta})\hat{Z}$, where $\hat{Z} = Z(\hat{\beta})$ and $Z(\beta) = D(\beta)\beta + e(\beta)$ (Jorgensen 1992). Thus, $\hat{\beta}$ can be regarded as the generalized least-squares estimate of the following linear model:

$$\hat{Z} = D(\hat{\beta})\beta + e \quad \text{and} \quad \mathrm{Var}(e) = \sigma^2 V(\hat{\theta})^{-1}. \tag{2.22}$$

We can extend the existing diagnostics for linear regression to Rician regression (Cook and Weisberg 1982; Jorgensen 1992; Wei 1998). Because $V(\hat{\theta})^{-1}$ reduces to an identity matrix at a high SNR, model (2.22) just reduces to a standard linear regression model.

We introduce two influence measures based on the representation of the linear model (2.22) as follows.

i) The residuals and standardized residuals are given by

$$\hat{r}_i = u_i^T \hat{V}(\hat{\theta})^{1/2}\{\hat{Z} - D(\hat{\beta})\hat{\beta}\} \quad \text{and} \quad \hat{t}_i = \sigma^{-1}\hat{r}_i/\sqrt{1 - h_{i,i}}, \tag{2.23}$$

where u_i is an $n \times 1$ vector with $i-$th element and all others zero, and where $\{h_{i,i} : i \leq n\}$ are the diagonal elements of the hat matrix H defined by

$$H = V(\hat{\theta})^{1/2}D(\hat{\beta})\left[D(\hat{\beta})^T V(\hat{\theta})D(\hat{\beta})\right]^{-1}D(\hat{\beta})^T V(\hat{\theta})^{1/2}. \tag{2.24}$$

Residuals are highly informative about the compatibility of a postulated model with the

observed data. If a Rician model is correct, residuals should be centered around zero, and plots of the residuals should exhibit no systematic tendencies. Exploring residual plots may reveal non-constant variance, curvature and the need for transformation in the regression, and therefore the analysis of residuals has been among the most widely used tools for assessing the validity of model specification (Cook and Weisberg 1982). To assess the magnitudes of the residuals, we compare the standardized residuals with the conventional benchmark 2.5. In other words, we regard the $i-$th data point (S_i, x_i) as having excess influence if $|\hat{t}_i|$ is larger than 2.5. We will plot the number of outliers at each voxel of the MR image. Voxels with many outliers need some further exploration.

ii) Cook's distance (Cook and Weisberg 1982) can be defined as

$$C_i = (\hat{\beta} - \hat{\beta}_{(i)})^T [D(\hat{\beta})^T V(\hat{\theta}) D(\hat{\beta})](\hat{\beta} - \hat{\beta}_{(i)})/\sigma^2, \tag{2.25}$$

where $\hat{\beta}_{(i)}$ denotes the maximum likelihood estimate of β based on a sample size of $n-1$ with the $i-$th case deleted. Instead of calculating $\hat{\beta}_{(i)}$ directly, we compute the first order approximation of $\hat{\beta}_{(i)}$, denoted by $\hat{\beta}^I_{(i)}$, which is given by

$$\hat{\beta}^I_{(i)} \approx \hat{\beta} - [D(\hat{\beta})^T V(\hat{\theta}) D(\hat{\beta})]^{-1} V_i(\hat{\theta})^{1/2} D_i(\hat{\beta}) \hat{r}_i/(1 - h_{i,i}),$$

where $D_i(\hat{\beta})^T$ is the $i-$th row of $D(\hat{\beta})$. Therefore, we get the first-order approximation of C_i, denoted by C_i^I, as $C_i^I = h_{i,i}\hat{t}_i^2/(1 - h_{i,i})$. Following Zhu and Zhang (2004), we compare nC_i^I with $3p$ to reveal the level of influence of (S_i, x_i) for each i at each voxel.

2.3.4 3D and 2D Graphics

We use 3D images of our various statistical measures to isolate all voxels in the image where specification of a Rician model is problematic. After computing the p-value of each test statistic (CM_1, CM_2, CK_1, or CK_2) at each voxel of the image, we create a 3D image of the $-\log_{10}(p)$ values for each statistic and then explore these values efficiently

across all voxels. In addition, we calculate t_i and C_i^I, compute the number of outliers at each voxel, and create a 3D image for each of these influence measures (Luo and Nichols 2003). For instance, if the p-value of CK_1 in a specific voxel is smaller than a given significance level, then we have strong evidence that the noise characteristics at that voxel are non-Rician and are likely to derive from non-physiological sources that may obscure valid statistical testing in those regions. Moreover, a large number of outliers appearing in several images taken sequentially, as they are in fMRI, may indicate a problematic noise source spanning the duration over which those images are obtained, as is often true of head motion, signal drift, and other similar artifacts. In addition, we also inspect the spatial clustering behavior of the voxels, which have large values of influence measures and test statistics, such as the cluster sizes of groups of outliers. More detailed examination of the 2D graphs for these voxels is indicated. These graphs include maps of the number of outliers pre slice and per image, index plots of influcence measures, and various plots of residuals that can reveal anomalies such as non-constant variance, curvature, transformations, and outliers in the data (Cook and Weisberg 1982; Luo and Nichols 2002). Thus, these 2D graphs of our diagnostic measures are used to help identify the nature and source of the disagreement between the Rician model and the observed MR signals at a particular voxel.

2.4 Simulation Studies

We conducted three sets of Monte Carlo simulations to examine the accuracy of using the Rician model, the two normal models and test statistics under differing experimental settings. The first set illustrated the performance of the Rician model and the two normal models for ADC imaging. The second set of simulations evaluated the sensitivity of the goodness-of-fit test statistics in detecting multiple tensor compartments within individual voxels of a DTI data set. The third set of Monte Carlo simulations evaluated the sensitivity of the goodness-of-fit statistics in detecting head motion in MR images.

2.4.1 Apparent diffusion coefficient mapping

The first set of Monte Carlo simulations was to compare the estimated ADC using the Rician model (2.2) and the two normal models (2.6) and (2.7). We set $d = 2 \times 10^{-3}$ mm^2/s, $S_0 = 500$, $b = [0, 50, 100, \cdots, 1100]$s/mm^2, and five different S_0/σ $\{2, 4, 6, 10, 15\}$ for all Monte Carlo simulations. For $S_0/\sigma = 2$, the values of the SNR were in the range of [0.366, 2]. At each S_0/σ, 4,000 diffusion weighted data sets were generated. Under each model, we calculated the parameter estimates $\hat{\theta} = (\hat{d}, \hat{S}_0, \hat{\sigma}^2)$. We finally calculated the biases, the empirical standard errors (SE), and the mean of the standard error estimates (SEE) based on the results from the 4,000 simulated ADC data sets (Table 2.1). At all S_0/σ, the estimates from model (2.2) had smaller biases, but larger SEs, whereas models (2.6) and (2.7) had larger biases, but smaller SEs. When $S_0/\sigma \geq 15$, models (2.2), (2.6) and (2.7) had comparable biases and SEs in the parameter estimates. In addition, the SE and its corresponding SEE are relatively close to each other when $S_0/\sigma \geq 4$.

2.4.2 Evaluating the test statistics for DTI data assuming the presence of fiber crossings

We assessed the empirical performance of CK_i and CM_i for $i = 1, 2$ as our test statistics for detecting the misspecified single diffusion model (2.8) when two diffusion compartments were actually present in the same voxel. Simulated data were drawn from the diffusion model (2.9) with 2 diffusion compartments, in which $p_1 = 1 - p_2$ was set at either 0.0 or 0.5, $D_1 = \mathrm{diag}(1.7, 0.2, 0.2)$ $(\times 10^{-3} mm^2/s)$, and $D_2 = \mathrm{diag}(0.2, 1.7, 0.2)$ $(\times 10^{-3} mm^2/s)$. In particular, $p_1 = 0.0$ corresponded to a single diffusion compartment, whereas $p_1 = 0.5$ corresponded to two diffusion compartments. The principal directions of D_1 and D_2 were, respectively, at $(1, 0, 0)$ and $(0, 1, 0)$. The mean diffusivity trace$(D)/3$ for both D_1 and D_2 was set equal to 1×10^{-3} mm^2/s, which is typical of values for normal cerebral tissue (Skare et al. 2000). We generated the Rician noise with $S_0 = 150$ and selected S_0/σ to be 5, 10, 15, 20, and 25, respectively. Our DTI scheme comprised

Table 2.1: ADC imaging: Bias and SD of three components of $\hat{\theta}$. TRUE denotes the true value of the regression parameters; BIAS denotes the bias of the mean of the regression estimates; SE denotes the empirical standard errors; SEE denotes the mean of the standard error estimates. Five different S_0/σ $\{2, 4, 6, 10, 15\}$ and 10,000 simulated datasets were used for each case.

		$R(\mu_i, \sigma^2)$			$N(\sqrt{\mu_i^2 + \sigma^2}, \sigma^2)$			$N(\mu_i, \sigma^2)$		
	S_0/σ	σ^2	S_0	d	σ^2	S_0	d	σ^2	S_0	d
TRUE	2	62500	500.00	2.000	62500	500.00	2.000	62500	500.00	2.000
BIAS	2	-13413	14.31	0.249	-23715	18.87	-0.749	-29683	15.73	-1.403
SE	2	19023	168.52	1.960	15494	139.40	1.241	10719	102.62	0.364
SEE	2	24123	255.12	2.460	16320	175.54	1.419	13009	88.91	0.378
TRUE	4	15625	500.00	2.000	15625	500.00	2.000	15625	500.00	2.000
BIAS	4	-1938	-5.46	0.080	-4542	-6.32	-0.284	-5014	-19.73	-0.711
SE	4	5218	82.05	0.909	3658	76.92	0.637	3488	65.95	0.332
SEE	4	6106	108.88	0.998	4285	79.48	0.611	4040	60.23	0.343
TRUE	6	6944	500.00	2.000	6944	500.00	2.000	6944	500.00	2.000
BIAS	6	-718	-2.26	0.016	-1680	-4.55	-0.127	-1746	-12.39	-0.371
SE	6	2409	51.99	0.469	1710	50.02	0.353	1702	65.95	0.332
SEE	6	2708	66.86	0.500	1998	55.36	0.392	1972	60.23	0.343
TRUE	10	2500	500.00	2.000	2500	500.00	2.000	2500	500.00	2.000
BIAS	10	-230	0.43	-0.025	-414	-1.08	-0.033	-422	-4.20	-0.138
SE	10	893	31.45	0.218	651	30.68	0.204	661	29.32	181.80
SEE	10	938	37.34	0.242	683	34.65	0.228	786	32.62	196.24
TRUE	15	1111	500.00	2.000	1111	500.00	2.000	1111	500.00	2.000
BIAS	15	-109	-0.23	0.008	-141	-0.60	-0.015	-143	-2.03	-0.065
SE	15	339	20.20	0.136	303	20.18	0.135	307	19.94	0.127
SEE	15	396	24.24	0.149	365	23.68	0.148	366	23.04	0.138

6 baselines, 30 diffusion weighted uniformly arranged directions at b_1, and the same set of gradient directions at b_2. We chose three combinations of (b_1, b_2): $(1000, 1000)$, $(1000, 3000)$, and $(3000, 3000)$ s/mm^2 in order to examine the sensitivity of differing b factors in detecting multiple fiber directions. For each simulation, 1,000 simulated data sets were used to estimate the nominal significance level (i.e., rejection levels for the null hypothesis). Finally, for each simulated data set, we applied the resampling method with $Q = 1000$ replications to calculate the four p-values of CK_i and CM_i for $i = 1, 2$ and then applied the false discovery rate procedure to correct for multiple comparisons at a significance level 5% as suggested by a reviewer.

Table 2.2 presents estimates for the rejection rates of the four test statistics after correction for multiple comparions based on the false discovery rate procedure. We observed that in a single compartment, the rejection rates of CK_i and CM_i for $i = 1, 2$ were smaller than the nominal level. Overall, the rejection rates in all cases were relatively accurate, and the Type I errors were not excessive. These findings suggested that the resampling method worked reasonably well under the null hypothesis. Differing (b_1, b_2) combinations strongly influenced the finite performance of the four test statistics in detecting the presence of two compartments. Specifically, compared with other (b_1, b_2) combinations, $(b_1, b_2) = (1000, 3000)$ s/mm^2 provided the best performance. Under $(b_1, b_2) = (1000, 3000)$ s/mm^2, CK_1 and CM_1 provided substantial power to detect the presence of two diffusion compartments. Compared with the other three statistics, CK_1 performed well; moreover, consistent with our expectations, increasing S_0/σ reduced the Type II errors and improved the power of the statistic CK_1 to detect the presence of two compartments. Therefore, these simulations suggested that the choice of b strongly influenced the performance of these test statistics and the test CK_1 was a useful tool for detecting the presence of multiple compartments. The selection of optimal b values in detecting multiple compartments warrants further research (Alexander et al. 2002; Jones et al. 1999).

Table 2.2: Comparison of the rejection rates for the test statistics CK_1, CM_1, CK_2, and CM_2 under the two-DT model, in which $f(x_i, \beta) = S_0[p_1 \exp(-b_i r_i^T D_1 r_i) + (1 - p_1) \exp(-b_i r_i^T D_2 r_i)]$ at a significance level of 0.05 after correction for multiple comparisons based on the false discovery rate. The first DT compartment is $D_1 = \text{diag}(1.7, 0.2, 0.2)$ and the second DT compartment is $D_2 = \text{diag}(0.2, 1.7, 0.2)$. Five different S_0/σ values $\{5, 10, 15, 20, 25\}$ and 1,000 simulated data sets were used for each case.

		(b_1, b_2) $\times 1000 s/mm^2$											
		$(1,1)$				$(1,3)$				$(3,3)$			
SNR	p_1	CK_1	CK_2	CM_1	CM_2	CK_1	CK_2	CM_1	CM_2	CK_1	CK_2	CM_1	CM_2
5	1	0.02	0.01	0.03	0.04	0.05	0.03	0.04	0.04	0.07	0.07	0.05	0.06
10	1	0.04	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.03	0.03	0.04	0.04
15	1	0.04	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.02	0.03	0.03	0.04
20	1	0.02	0.02	0.03	0.04	0.03	0.04	0.03	0.04	0.02	0.03	0.03	0.04
25	1	0.01	0.02	0.02	0.02	0.04	0.03	0.05	0.04	0.02	0.02	0.026	0.02
5	1	0.01	0.02	0.03	0.03	0.05	0.05	0.08	0.07	0.08	0.09	0.05	0.06
10	1	0.05	0.04	0.02	0.02	0.23	0.08	0.22	0.12	0.04	0.02	0.01	0.02
15	1	0.09	0.05	0.02	0.02	0.43	0.11	0.39	0.15	0.08	0.01	0.01	0.01
20	1	0.16	0.09	0.03	0.03	0.61	0.11	0.59	0.22	0.09	0	0	0
25	1	0.26	0.18	0.02	0.02	0.75	0.14	0.71	0.19	0.12	0	0	0

2.4.3 Evaluating the test statistics in the presence of head motion

We also assessed the empirical performances of CK_i and CM_i for $i = 1, 2$ as test statistics for detecting the misspecified single diffusion model (2.8) at a single voxel in the presence of head motion. We simulated data contaminating head motion in the image as follows. We used a DTI scheme starting with 5 baselines and followed with 45 diffusion weighted uniformly arranged directions at $b_1 = 1000 s/mm^2$. We simulated data from the diffusion model (2.8) with $D_1 = \text{diag}(0.2, 1.7, 0.2)$ $(\times 10^{-3} mm^2/s)$ in the first $[50 \times p_1]$ acquisitions, and then generated data from the diffusion model (2.8) with $D_2 = \text{diag}(0.7, 0.7, 0.7)$ $(\times 10^{-3} mm^2/s)$ from the last $50 - [50 \times p_1]$ acquisitions, where $[\cdot]$ denoted the largest integer smaller than $50 \times p_1$. In addition, the probability p_1 was selected to be 0.5 and 0.7, which reflected the different degrees of head motion. We also generated Rician noise

Table 2.3: Comparison of the rejection rates for the test statistics CK_1, CK_2, CM_1, and CM_2, under the presence of head motion at a significance level of 0.05 after correction for multiple comparisons based on the false discovery rate. The first $[50 \times p_1]$ acquisitions were generated from a single diffusion model with $D_1 = \text{diag}(0.2, 1.7, 0.2)$ and the last $50 - [50 \times p_1]$ acquisitions were generated from a single diffusion model with $D_2 = \text{diag}(0.7, 0.7, 0.7)$. Five different S_0/σ values $\{5, 10, 15, 20, 25\}$ and 1,000 simulated data sets were used for each case.

	p_1							
	0.7				0.5			
SNR	CK_1	CK_2	CM_1	CM_2	CK_1	CK_2	CM_1	CM_2
5	0.02	0.02	0.05	0.05	0.03	0.01	0.05	0.05
10	0.04	0.08	0.09	0.16	0.07	0.06	0.11	0.16
15	0.09	0.14	0.10	0.23	0.08	0.09	0.10	0.23
20	0.11	0.19	0.09	0.31	0.12	0.13	0.12	0.31
25	0.12	0.23	0.08	0.32	0.13	0.13	0.10	0.31

from (2.1) with $S_0 = 150$ and set S_0/σ to be 5, 10, 15, 20, and 25 respectively. For each simulation, 1,000 simulated data sets were used to estimate the nominal significance level (i.e., rejection levels for the null hypothesis). Finally, for each simulated data set, we applied the resampling method with $Q = 1000$ replications to calculate the four p-values of CK_i and CM_i for $i = 1, 2$ and then applied the false discovery rate procedure to correct for multiple comparisons at a significance level of 5% as suggested by a reviewer.

Table 2.3 presented estimates for the rejection rates of our four statistics after correction for multiple comparions based on the false discovery rate procedure. Compared with the other three statistics, CM_2 was the most sensitive statistic in detecting head motion. Moreover, consistent with our expectations, increasing S_0/σ reduced the Type II errors and improved the power of the statistic CM_2 for detecting the presence of two compartments. However, the other three statistics CK_1, CM_1, and CK_2 were not particularly sensitive in detecting head motion.

2.4.4 Diffusion Weighted Images with Head Motion

We acquired DWIs of the brain of a healthy adult male subject (right-handed; age 34 years). The imaging acquisition scheme $\{(b_i, r_i) : i = 1, \cdots, 38\}$ consisted of 3 baseline images with $b = 0$ s/mm^2 and 35 directions of diffusion gradients that were arranged uniformly in the 3-dimensional space at $b = 1000$ s/mm^2 (Hardin, Sloane and Smith 1994). Each DWI contained $256 \times 256 \times 65$ voxels. The subject was instructed to move his head deliberately during acquisition of images from the 28th to the 38th direction. Head motion varied from 2- to 6-degrees of rotation and 0- to 10- millimetres of translation, causing the diffusion weighted images to be moderately misaligned.

We used the Rician DTI model (2.8) for this analysis. We subsequently calculated at each voxel the ML estimate $(\hat{D}, \hat{S}_0, \hat{\sigma})$, three eigenvalue-eigenvector pairs of $\hat{D}$, denoted by $\{(m_i, e_i) : i = 1, 2, 3\}$, and the invariant measures including CL $= (m_1 - m_2)/M_1$, CP $= 2(m_2 - m_3)/M_1$, RA $= \sqrt{1 - 3M_2 M_1^{-2}}$, and FA $= \sqrt{1 - M_2(M_1^2 - 2M_2)^{-1}}$, where $m_1 \geq m_2 \geq m_3$, $M_1 = \mathrm{tr}(\hat{D})$, $M_2 = m_1 m_2 + m_1 m_3 + m_2 m_3$, and $M_3 = m_1 m_2 m_3$. We also calculated three test statistics $T_a =$ FA, $T_b = S(\hat{D}) + W(\hat{D})^{1.5}$, and $T_c = S(\hat{D}) - W(\hat{D})^{1.5}$, and their associated p-values, where $S(\hat{D}) = (M_1/3)^3 - M_1 M_2/6 + M_3/2$ and $W(\hat{D}) = (M_1/3)^2 - M_2/3$. We further set the significance level at 1% and used the p-values of T_a, T_b, and T_c to classify the morphology of the tensor at each voxel (Zhu, Xu, Amir, Hao, Zhang, Alayar, Ravi, and Peterson 2006).

We then assessed the quality of these diffusion weighted images using our diagnostic methods. We searched for artifacts, scanner instability problems, and voxels that contained outliers; in addition, we obtained diagnostic measures, generated scan summaries, and applied graphical tools. We estimated the p-values of the four test statistics CK_1, CK_2, CM_1, and CM_2 using the resampling method in Section 2.3 of this paper.

We plotted maps of scan summaries to identify possible artifacts and acquisition problems in the DW images. Translational and rotational parameters (Fig. 2.3), obtained from FLIRT in FSL (http://www.fmrib.ox.ac.uk), detected rightward rotation of

2 to 6 degrees and 0 to 10 mm translation beginning in the 28th acquisition (Jenkinson and Smith 2001; Jenkinson, Bannister, and Smith 2002). Outlier statistics detected these head motions as well. The outlier count per slice and per direction showed clearly that a large batch of outliers appeared in almost all of the slices along the last ten directions (red to white on the color spectrum in Fig. 2.4). Furthermore, we performed a spatial independent component analysis (ICA) on the 16 slices covering the middle part of each directional DWI (baseline images excluded). Using the Bayesian Information Criterion (BIC), we selected 8 independent components and plotted the associated time series from the spatial ICA. The time series associated with the 4th, 7th, and 8th components revealed the deliberate rotation and translation from the 28th to 33rd acquisitions. The detailed information about the ICA results can be found in the supplementary report.

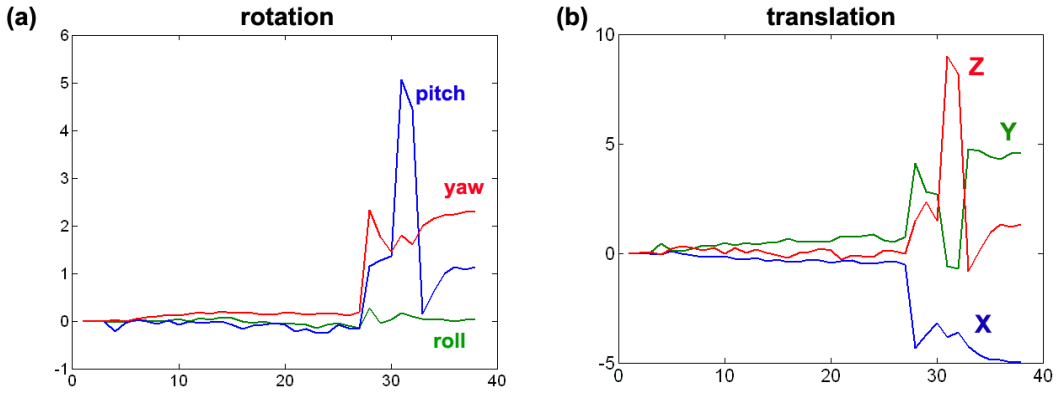

Figure 2.3: Scan summaries for a set of DWIs from a single subject: (a) translational parameters; (b) rotational parameters.

To reduce or eliminate motion artifacts, we used the rigid-body transformation method to co-register all other DW images to the first DW image while properly reorienting the diffusion gradients (Rohde, Barnett, Basser, Marenco, and Pierpaoli 2004). Particularly, we applied the translational and rotational parameters obtained from FLIRT and used a 7th order interpolation method to resample the DW images. After coregistration, new translational and rotational parameters (not shown here) revealed that the DW images were properly aligned. We then assessed the realigned DW images using

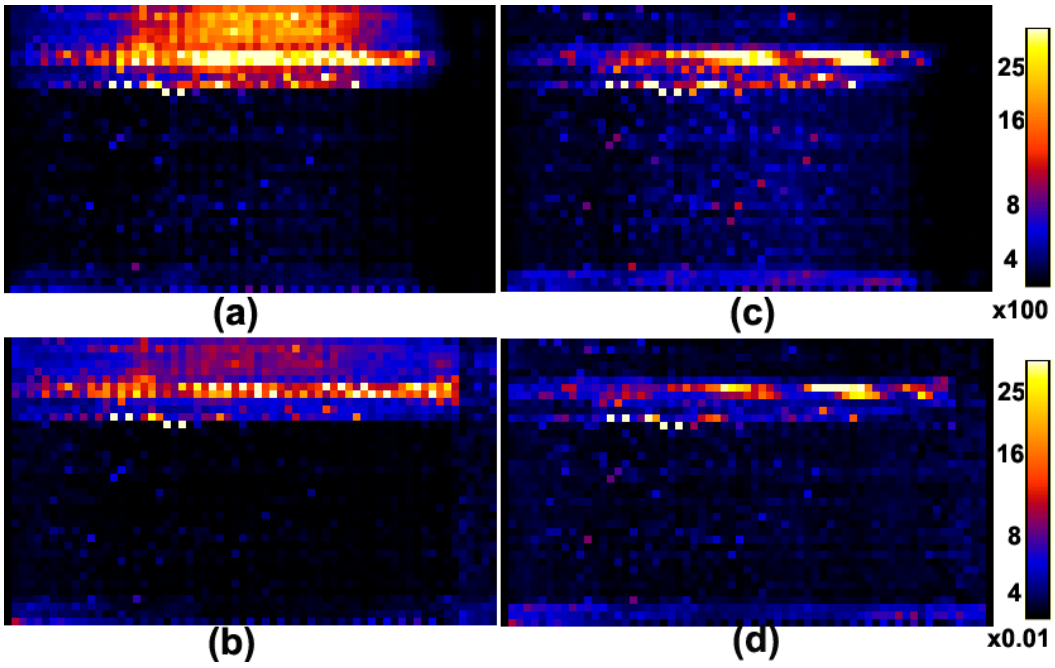

Figure 2.4: Assessing the effect of applying a coregistration algorithm to diffusion weighted images from a single subject: outlier count per slice and per direction (a) before coregistration and (c) after coregistration; percentages of outliers per slice and per direction (b) before coregistration and (d) after coregistration.

our diagnostic procedure and used the Rician model (2.8) to process the reoriented DW images.

Our diagnostic procedure can be used to quantify the efficacy of the coregistration and reslicing algorithms, and to identify potential problems that remain in the DW images after registration and reslicing. We observed a substantial decline in the number of outlier counts per slice and per direction compared with the non-realigned images, as well as a decline in the percentage of outliers per slice and per direction after coregistration (Figs. 2.4 a, b, c, and d). Furthermore, we examined voxels having 0-10 outliers and found that motion correction using coregistration significantly decreased the percentage of voxels having 4-10 outliers from 2.85% to 1.41%. However, despite the efficacy of this method for correcting motion artifacts, 5.7% of the voxels still contained at least three outliers after coregistration, and the 28th to 33rd acquisitions (red to white on the color spectrum) contained a number of outliers (Fig. 2.4c, red to white on the color

spectrum). This may indicate that the rigid-body transformation and the interpolation method cannot completely remove the effect of moderate and large head motions in MRIs.

The 3D images of the $-\log_{10}(p)$ values for the test statistics CK_1, CK_2, CM_1, and CM_2 were more sensitive and specific in assessing the quality of the DW images (Figs. 2.6). A p-value of 0.001 corresponded to a $-\log_{10}(p)$ value of 3; thus a voxel having a $-\log_{10}(p)$ value greater than or equal to 3.0 was conventionally regarded as statistically significant and in need of further investigation. In all maps of $-\log_{10}(p)$ values of the test statistics, we focused on voxels having significant p-values (white) and then searched for systematic patterns of these voxels in the brain. We found several notable changes after coregistration as follows. The number of voxels having large $-\log_{10}(p)$ values for the CK_1, CK_2, CM_1, and CM_2 statistics declined dramatically following coregistration (Fig. 2.6). We also used the resampling methods in Section 2.3 to calculate the corrected $-\log_{10}(p)$ values, but no significant voxel was detected for all four test statistics at the 5% significance level before and after coregistration. Moreover, compared with CK_1 and CK_2, CM_1 and CM_2 were more sensitive measures for detecting head motion.

Assessing the quality of DW images was crucial for further processing images. As shown above (Fig. 2.6), the maps of the $-\log_{10}(p)$ values of the test statistics not only provided detailed information about the goodness of fit of the fitted Rician model with the DW images (Fig. 2.6), but also these maps indicated possible artifacts existing in the DWIs. Those artifacts strongly influenced the estimation of the DTs, the classification of tensor morphologies, the reconstruction of fiber tracts, and the quantification of uncertainty in tensor estimation and tractography. Therefore, we also assessed the prevalence of the four morphological classes of DTs (nondegenerate, oblate, prolate and isotropic) in a single slice before and after coregistration. Before coregistration, we found that 59.97% were isotropic, 9.37% were oblate, 23.06% were prolate and 7.61% were nondegenerate. Following coregistration, we found that 48.09% were isotropic, 11.35% were

oblate, 28.11% were prolate and 12.45% were nondegenerate. Most tractography algorithms can only track fibers across voxels containing either nondegenerate or prolate DTs, which accounted for 40.56% of the total number of voxels on this slice after coregistration, compared with 30.67% before coregistration. Moreover, we also found moderate discrepancy between the estimated principal directions before and after coregistration (not presented here).

To assess these DW images before and after coregistration, we also examined 3D images of standardized residuals and Cook's distances. Specifically, we searched the standardized residuals (or Cook's distance) in all voxels across all slices and directions to identify voxels having large numbers of positive and negative outliers (i.e., data points of excessive influence). For illustration, we compared the standardized residuals at the 30th slice from the 32nd acquisition before and after coregistration (Fig. 2.7). Before coregistration, this slice contained many positive and negative residuals (Figs. 2.6a and 2.6b). After coregistration, the number of positive and negative residuals dramatically declined (Figs. 2.6c and 2.6d). However, even after coregistration, some motion artifacts or other unspecified problems remained in the resliced DW images. Developing methods for identifying the precise sources of non-Rician noise and correcting for them in the resliced DW images will require further research.

For voxels having either many outliers or substantial misspecification of the Rician model, we examined multiple 2D graphs to try to identify the causes of the outliers and of model misspecifications. To illustrate this process, we considered the data at a single voxel (at location (100, 69, 30)) before coregistration. The p-values for CK_1, CK_2, CM_1, and CM_2 were 0.21, 0.13, 0.03, and 0.01, respectively. The index plots of the standardized residuals and Cook's distances (Fig. 2.8a, 2.8b) revealed that the 4th, 8th, and 34th observations were likely outliers. A plot of the standardized residuals against the raw MRI values (Fig. 2.8c) revealed a strong linear relationship between residuals and the raw MRI values (Cook and Weisberg 1982). Furthermore, we observed

a nonlinear relationship (Fig. 2.8d) of Cook's distances against raw MRI values. Together these plots (Fig. 2.8c, 2.7d) indicated that a Rician model (2.8) did not fit the MRI data satisfactorily. Further improvements in model specification or post-acquisition processing are needed to identify and address the non-Rician sources of noise in the images.

Our diagnostic procedure effectively identified head motion artifacts in DW images. Coregistration improved image quality, but substantial non-stochastic noise sources remained in the 28th to 33rd acquisitions. One solution is to remove these slices from the subsequent analysis; alternatively, we may resort to a robust estimate of DTs to reduce the deleterious statistical effects of these outliers. The 3D images of the test statistics further detected additional physiological noise, such as cardiac pulsation, in DW images. Additional 2D statistical maps may identify the causes of statistically significant voxels and the location of outliers.

2.4.5 Concluding Remarks

We have developed estimation algorithms for fitting a Rician regression model and the associated two normal models, and proposed a diagnostic procedure for systematically assessing the quality of MR images at all levels of the SNR. The key features of our procedures include: calculating test statistics that assess the validity of the assumptions of the statistical models for stochastic noise in MR images; use of influence measures to identify artifacts and problems with image acquisition; and multiple graphical tools for visual evaluation of the appropriateness of the model assumptions. Simulations showed the effectiveness of our test statistics in detecting the presence of multiple compartments. Moreover, an in-vivo study demonstrated the effectiveness of our procedures in locating voxels that contain unreliable data due to motion artifacts or to problems with imaging acquisition. Our findings suggest that our approach to assessing the quality of MR images is both rigorous and computationally practical.

Our diagnostic procedure differs substantially from previous model-free methods,

such as independent-component analysis and motion correction algorithms, for detecting noise components in MRI. Most of those model-free methods cannot be used to detect non-stochastic noise components at the voxel level, since they can only provide information about MRI at the whole brain level. In addition, some of those model-free methods are limited to a specific imaging modality. For instance, although an independent-component analysis (ICA) method was recently proposed to identify independent componts (ICs) associated with task-related motion, and then discard those ICs in order to reduce motion effects on realigned fMRI data (Kochiyama, Morita, Okada, Yonekura, Matsumura, Sadato 2005), this ICA method cannot be directly applied to other imaging modalities, such as DWI. Particularly, for DWI, we cannot discard the ICs corresponding to head motion without changing the gradient directions, which requires further research. In contrast, as shown in Section 2.4.4, our diagnostic procedure is a model-based method that uses goodness-of-fit statistics and diagnostic measures to systematically detect non-stochastic noise components at each voxel of the MRI data. Subsequently, our diagnostic procedure can combine the information from all voxels of the brain volume to identify large non-stochastic noise sources, such as head motion at the whole volume level.

Our procedure takes a further step by studying how to use existing information in the MRI data to check model assumptions and to identify imaging artifacts that may undermine applications or interpretations of the MR images. Our diagnostic procedure can also be applied to systematically check the MRI data even after these MRI data have been processed by existing noise removal methods such as rigid-motion correction and ICA. Moreover, our diagnostic procedure can be used to detect the presence of the partial volume effect, whereas those existing methods, such as the motion correction method, cannot. Nevertheless, our procedure assesses the quality of MRI statistically and cannot replace various preprocessing techniques, such as registration and smoothing methods.

2.5 Appendix

The following assumptions are needed to facilitate the technical details, although they are not the weakest possible conditions.

(C1) $(S_i, \mathbf{x}_i)$ are independently and identically distributed and the conditional distribution of S_i given $\mathbf{x}_i$, denoted by $p(S_i|\mathbf{x}_i, \theta)$, follows the Rician distribution $R(\mu_i(\beta), \sigma^2)$ with $\mu_i(\beta) = f(\mathbf{x}_i, \beta)$.

(C2) The true value $\theta_* = (\beta_*, \sigma_*^2)$ is unique and an interior point of $\Theta = \mathcal{B} \times [a_0, b_0]$, where $\mathcal{B}$ is a compact set in R^p and $\infty > b_0 > a_0 > 0$.

(C3) $f(\mathbf{x}, \beta)$ is twice continuously differentiable with respect to β and $|f(\mathbf{x}, \beta)|$, $||\partial_\beta f(\mathbf{x}, \beta)||$, and $||\partial_\beta^2 f(\mathbf{x}, \beta)||$ are bounded by some integrable function $F_0(\mathbf{x})$ with $E[F_0(\mathbf{x})^4] < \infty$, where $\partial_\beta = \partial/\partial\beta$.

(C4) $f(\mathbf{x}, \beta)$ is identifiable, that is $f(\mathbf{x}, \beta_1) = f(\mathbf{x}, \beta_2)$ for all $\mathbf{x}$ ensures that $\beta_1 = \beta_2$.

(C5) The elements of $I(\theta) = \int \{\partial_\theta \log\ p(S|\mathbf{x}, \theta)\}^{\otimes 2} p(S, \mathbf{x}) dS d\mathbf{x}$ are continuous in θ and is nonsingular at $\theta = \theta_*$, where $\partial_\theta = \partial/\partial\theta$ and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for any vector $\mathbf{a}$.

2.5.1 Proof of Theorem 1

The proof of Theorem 1 (i) consists of two steps as follows:

Step 1. $\hat{\theta}$ is a consistent estimate of θ_* and

$$\sqrt{n}(\hat{\theta} - \theta_*) = n^{-1/2} \sum_{i=1}^{n} \psi(S_i, \mathbf{x}_i; \theta_*) + o_p(1). \tag{2.26}$$

Step 2. $(T_1(\cdot; \theta_*), \sqrt{n}(\hat{\theta} - \theta_*)^T)^T$ converges to the Gaussian process as described in Theorem 1 (i).

In Step 1, we primarily prove that $\hat{\theta}$ is a consistent estimate of θ_*, because assumptions (C3) and (C5) and the consistency of $\hat{\theta}$ ensure (2.26) (Theorem 5.39 in van der Vaart,

1998; p.65). Let

$$M_n(\theta) = n^{-1} \sum_{i=1}^{n} \log \frac{p(S_i, \theta)}{p(S_i, \theta_*)} \quad \text{and} \quad M(\theta) = E\{\log \frac{p(S_i, \theta)}{p(S_i, \theta_*)}\}.$$

To show $\hat{\theta}$ is a consistent estimate of θ_*, we check two sufficient conditions (Theorem 5.7 in ver der Varrt, 1998; p.45) as follows:

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \to^p 0 \quad \text{and} \quad \sup_{\theta : ||\theta - \theta_*|| \geq \epsilon} M(\theta) < M(\theta_*).$$

Because $\log p(S|\mathbf{x}, \theta)$ is Lipschitz in θ, $\{\log p(S|\mathbf{x}; \theta) : \theta \in \Theta\}$ is Glinvenko Cantelli (van der Vaart and Wellner, 1996; p. 122). Thus, $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|$ converges to zero almost surely.

To show $\sup_{\theta : ||\theta - \theta_*|| \geq \epsilon} M(\theta) < M(\theta_*)$, we check that $\log p(S_i; \beta, \tau)$ is identifiable (Lemma 5.35 in van der Vaart, 1998). Suppose that

$$\begin{aligned} G(S, \mathbf{x}) &= \log p(S|\mathbf{x}, \beta_1, \tau_1) - \log p(S_i|\mathbf{x}, \beta_2, \tau_2) \qquad (2.27) \\ &= S^2(\tau_2 - \tau_1) + \mu_2^2 \tau_2 - \mu_1^2 \tau_1 + 2 \log I_0(\mu_1 \tau S) - 2 \log I_0(\mu_2 \tau_2 S), \end{aligned}$$

where $\tau_1 = \sigma_1^{-2}$, $\tau_2 = \sigma_2^{-2}$, $\mu_1 = f(\mathbf{x}, \beta_1)$, and $\mu_2 = f(\mathbf{x}, \beta_2)$. We want to show that if $G(S|\mathbf{x}) = 0$ holds for all $(S, \mathbf{x})$, then (β_1, τ_1) must equal (β_2, τ_2). If $G(S, \mathbf{x}) = 0$, then $G(0, \mathbf{x}) = 0$ and $\partial_S G(S, \mathbf{x})|_{S=0} = 0$ hold. Thus, it can be shown that $G(0, \mathbf{x}) = \mu_2^2 \tau - \mu_1^2 \tau_1 = 0$ and $\tau_1 \mu_1 = \tau_2 \mu_2$. Thus, $f(\mathbf{x}, \beta_2) = f(\mathbf{x}, \beta_1)$ and $\tau_2 = \tau_1$. Assumption (C4) ensures that β_1 must equal β_2. Therefore, we can conclude that $\hat{\theta}$ converges to θ_* in probability and (2.26) holds.

In Step 2, we first show that the marginals of $(T_1(\cdot; \theta_*), \sqrt{n}(\hat{\theta} - \theta_*)^T)^T$ converge weakly to the corresponding marginals of the zero-mean Gaussian process $(G_1(\cdot; \theta_*), \nu_1)$. Based on Step 1, we only need to show $T_1(u; \theta_*)$ converges weakly to the marginal of Gaussian

process, $G_1(u;\theta_*)$. Because $|W_i| = |I_1(\mu_i S_i/\sigma^2)/I_0(\mu_i S_i/\sigma^2)| \leq 1$, we have

$$\text{Var}\{\mathbf{1}(\mathbf{x}_i^T\beta_* \leq u)[W_i(\theta_*)S_i - \mu_i(\mathbf{x}_i,\beta_*)]\} \leq \text{Var}\{W_i(\theta_*)S_i\} \leq \text{Var}(S_i) < \infty.$$

Therefore, the standard central limit theorem ensures that $T_1(u;\theta_*)$ converges to $G_1(u;\theta_*)$.

Second, we consider a class of measurable functions $\mathcal{F} = \{f(u) = [W(\theta_*)S - \mu(\mathbf{x},\beta_*)]\mathbf{1}(\mathbf{x}^T\beta_* \leq u), u \in [-\infty,\infty]\}$. First, $\mathcal{F}$ is a Vapnik and Cervonenkis (VC) class, which satisfies uniform entropy condition (van der Vaart and Weller, 1996; Sections 2.5 and 2.6). Based on Theorem 2.5.2 in ver der Vaart and Wellner (1996), to ensure that $\mathcal{F}$ is P-Donsker, we need to show: (a) $\mathcal{F}_\delta = \{f(u_1) - f(u_2) : f(u_1), f(u_2) \in \mathcal{F}, \int[f(u_1) - f(u_2)]^2 p(S,\mathbf{x})dSd\mathbf{x} < \delta\}$ and F_∞^2 are P-measurable for every $\delta > 0$. (b) $E[F(u)]^2 < \infty$, where $F(u)$ is the envelop function of $\mathcal{F}$. We can show that $\sup_{u_1,u_2} |\sum_{i=1}^n e_i[W_i(\theta_*)S_i - \mu(\mathbf{x}_i,\beta_*)]\mathbf{1}(u_2 \leq \mathbf{x}_i^T\beta_* \leq u_1)|$ is measurable for every n and every vector $(e_1,...,e_n) \in \{-1,+1\}^n$, because $\mathbf{1}(u_2 \leq \mathbf{x}_i^T\beta_* \leq u_1)$ is a measurable function and $|\sum_{i=1}^n e_i[W_i(\theta_*)S_i - \mu(\mathbf{x}_i,\beta_*)]\mathbf{1}(u_2 \leq \mathbf{x}_i^T\beta_* \leq u_1)|$ can only take 2^n possible values. Similarly, we also can show that F_∞^2 is P-measurable. Furthermore, because $F(u)$ can be chosen to be $S_i + f(\mathbf{x},\beta_*)$, assumption (C3) ensures that $E[F(u)^2] < \infty$. Finally, $\mathcal{F}$ is P-Donsker.

(ii) Applying the continuous mapping theorem yields Theorem 1 (ii).

(iii) The proof of Theorem (iii) is similar to that of Theorem (i), so details are omitted here.

(iv) Applying the continuous mapping theorem yields Theorem 1 (iv).

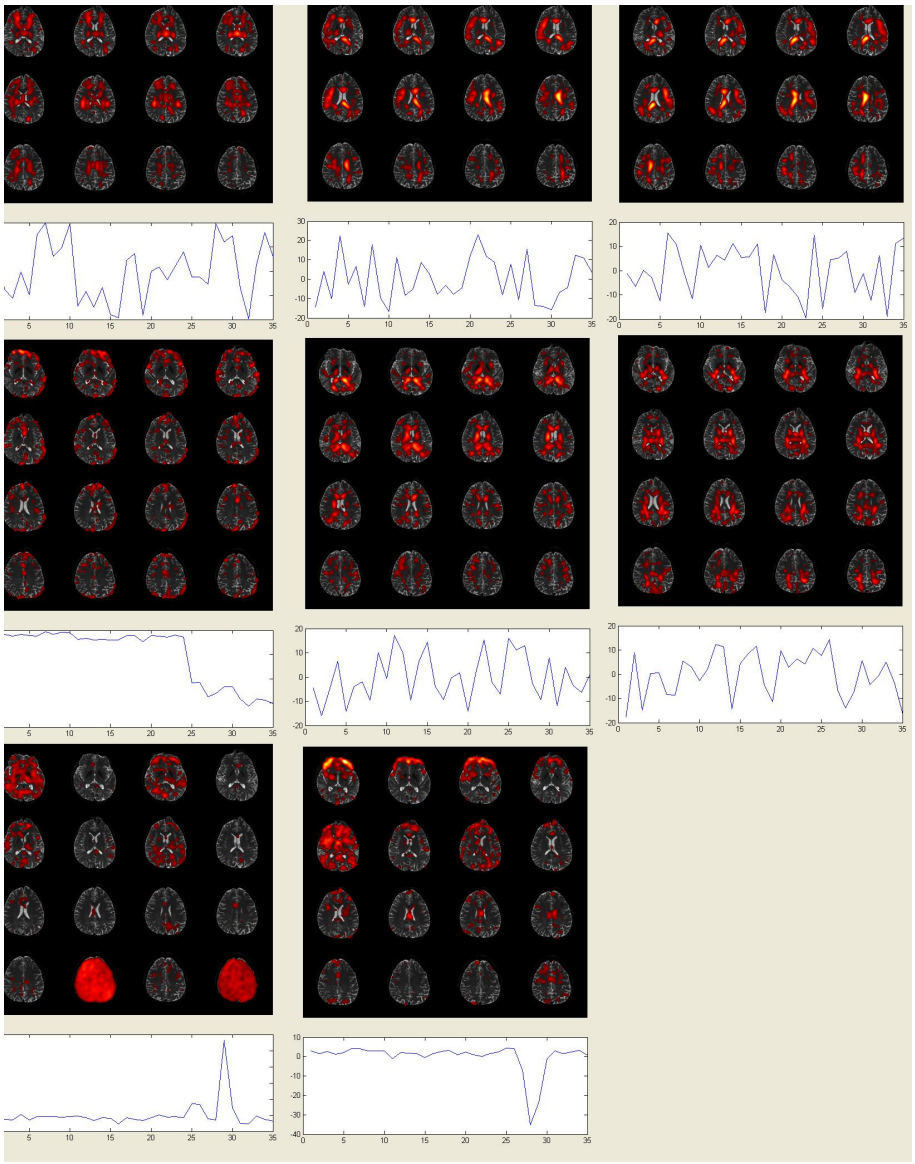

Figure 2.5: Maps of the eight selected independent components and their associated time series from a single subject. The 4th, 7th and 8th independent components are associated with the delibrate head motion.

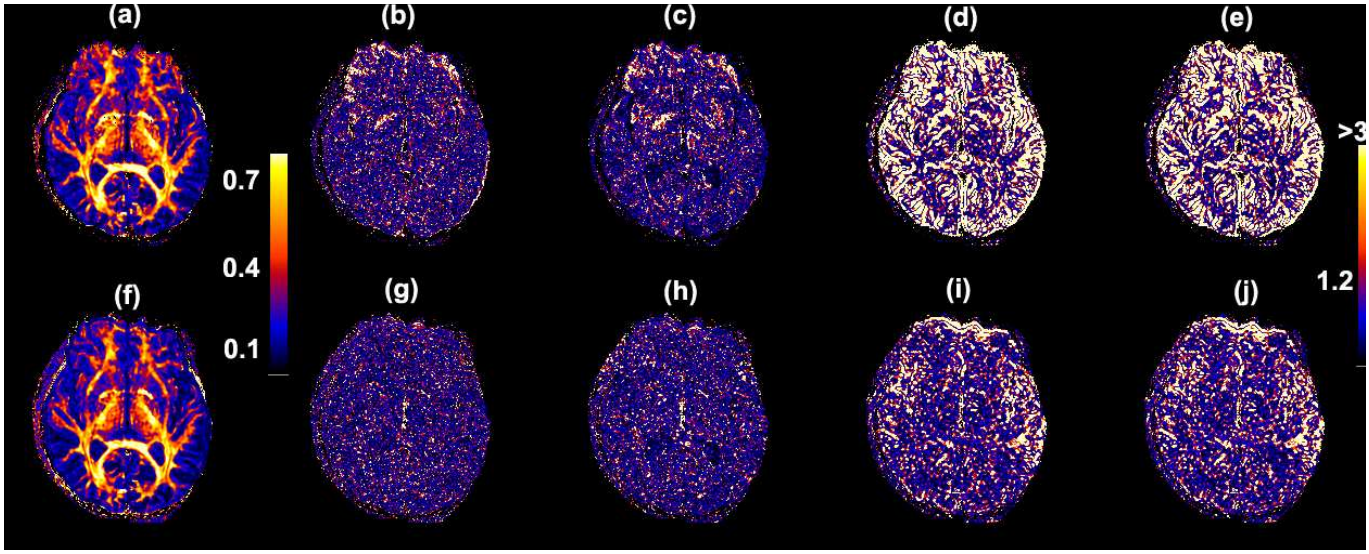

Figure 2.6: Maps of 3D images before coregistration (a-e) and after coregistration (f-j) in a single slice from a single subject. Before coregistration: (a) FA value; (b) $-\log_{10}(p)$ values of CK_1; (c) $-\log_{10}(p)$ values of CK_2; (d) $-\log_{10}(p)$ values of CM_1; (e) $-\log_{10}(p)$ values of CM_2. After coregistration: (f) FA value; (g) $-\log_{10}(p)$ values of CK_1; (h) $-\log_{10}(p)$ values of CK_2; (i) $-\log_{10}(p)$ values of CM_1; (j) $-\log_{10}(p)$ values of CM_2.

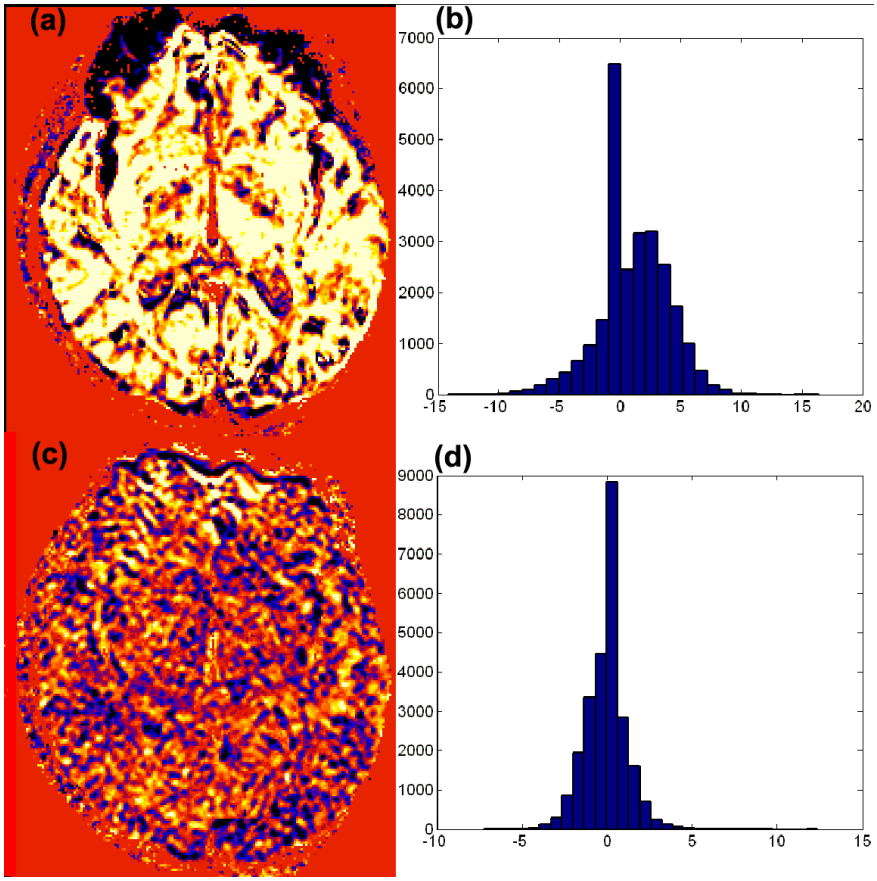

Figure 2.7: Plots of standardized residuals at the 30th slice of the 32nd acquisition before and after coregistration from a single subject: standardized residuals (a) before coregistration and (c) after coregistration; histgrams of standardized residuals (b) before coregistration and (d) after coregistration. Voxels in the black-to-blue range have large negative standardized residuals (< -2.5), while yellow to white voxels have large positive standardized residuals (> 2.5).

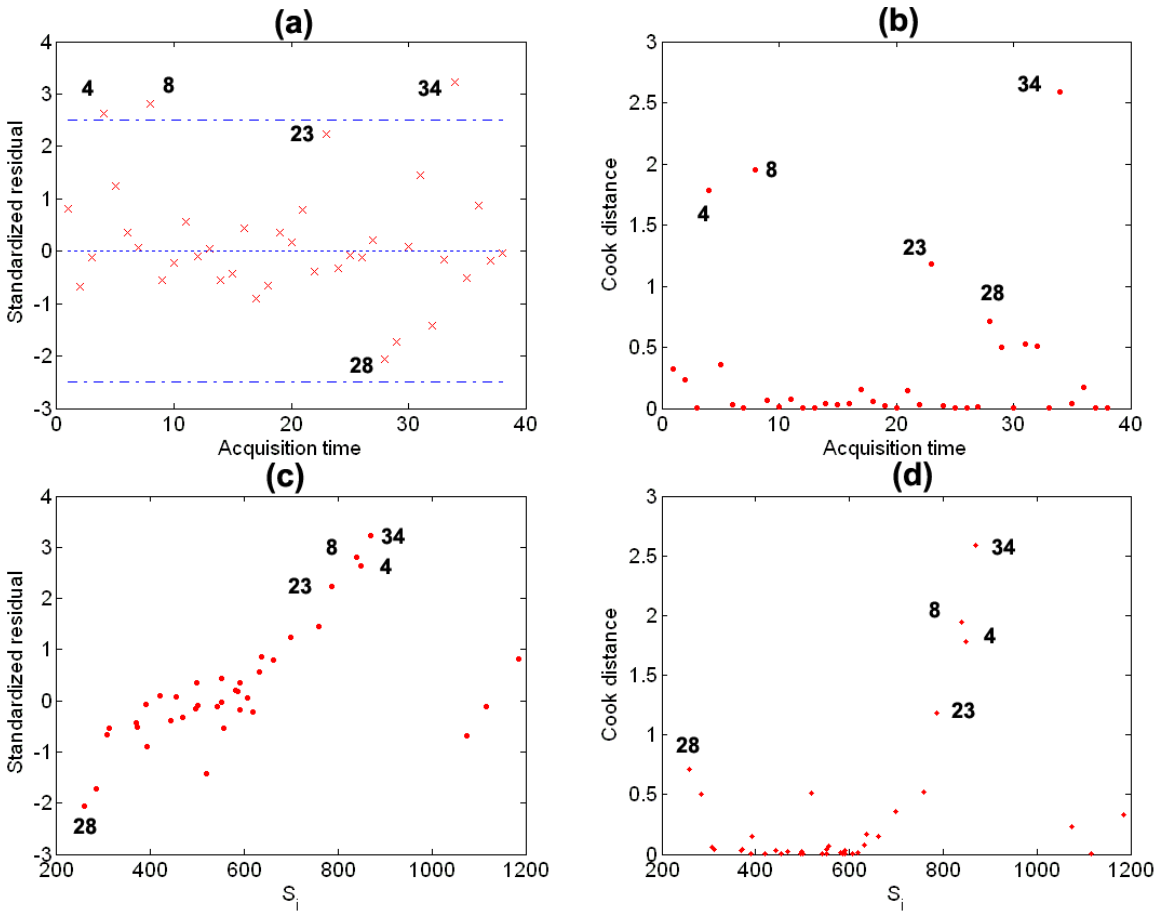

Figure 2.8: Multiple 2D graphs for a selected voxel (110, 69, 30) before coregistration from a single subject: (b) index plot of standardized residuals; (b) index plot of Cook's distances; (c) standardized residuals against raw data; (d) Cook's distances against raw data.

Chapter 3

Multiscale Adaptive Regression Models for Neuroimaging Data

3.1 Introduction

Magnetic resonance imaging (MRI) is an important medical imaging technique commonly used for understanding the neural development of neuropsychiatric and neurodegenerative disorders, and normal brains. For instance, by using anatomical MRI, statistical shape modeling and analysis have emerged as important tools for understanding neuroanatomical differences in cortical and subcortical structures (e.g., hippocampus) of the human brain *in vivo* across different populations or time (Thompson and Toga, 2002; Styner *et al.*, 2005). Diffusion tensor imaging can quantitatively assess the integrity of anatomical connectivity in white matter in the human brain *in vivo* (Basser *et al.*, 1994a, b; Schwartzman, 2005; Zhu *et al.*, 2007b). Functional MRI (fMRI) is a type of MRI scan for measuring the haemodynamic response relative to specific stimuli and behavioral tasks and has been widely used to understand functional integration of different brain regions (Friston, 2007; Huettel et al., 2004).

The goal of this article is to develop and apply a multiscale adaptive regression model (MARM) for the spatial and adaptive analysis of neuroimaging data and then to

demonstrate its superiority over the voxel-wise approach using simulated and real imaging data. MARM has three unique features: being spatial, being hierarchical and being adaptive. MARM builds a sphere with a given radius at all voxels, and then uses these consecutively overlapping spheres to capture local and global spatial dependence among different voxels. Thus, 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 neighborhood around each voxel and utilizes information in each of the nested spheres across all voxels. Finally, MARM combine all observations with adaptive weights in the voxels within the sphere of the current voxel to adaptively calculate parameter estimate and test statistic. Due to its hierarchical and adaptive natures, MARM slightly increases the amount of computational time in computing parameter estimate and testing statistic compared with the voxel-wise approach. Due to its spatial and adaptive features, MARM can efficiently utilize available information in the neighboring voxles to increase the precision of parameter estimates and the power of test statistics in detecting the subtle changes of brain structure and function.

MARM represents a novel generalization of the propagation−separation (PS) approach, which was originally developed for nonparametric estimation of regression curves or surfaces (Polzehl and Spokoiny, 2000, 2003, 2006). The PS approach has been used to smooth the image of parameter estimates and its associated image of standard errors, which are the output of the voxel-wise approach (Tabelow *et al.*, 2006). Recently, Tabelow *et al* (2008) borrowed the original PS idea and developed a multiscale adaptive linear model to adaptively and spatially denoise diffusion tensor images. Compared with the Gaussian distribution assumption of the multiscale adaptive linear model (Tabelow *et al.*, 2006), MARM is solely based on the pseudo-likelihood function, which is very desirable for the analysis of real neuroimaging data, since the distribution of the univariate (or multivariate) neuroimaging measurements often deviates from the Gaussian distribution (Ashburner and Friston, 2000; Luo and Nichols, 2003; Zhu *et al.*, 2007a).

More importantly, MARM provides a new probability framework for carrying out statistical inference on neuroimaging data from multiple subjects. The adaptive weights in the MARM differ from those in the PS approach and the covariance estimate of the adaptive estimator in MARM has a simpler form compared with that in Tabelow *et al.*, (2006).

According to the best of our knowledge, it is the first time that we establish the asymptotic properties of the parameter estimates and test statistics for MARM under some mild conditions. The existing theory for the PS approach was established for a class of nonparametric models based on exponential families under the propagation condition (Polzehl and Spokoiny, 2006). Since MARM is developed for neuroimaging data from multiple subjects, the theory of PS does not yield the asymptotic distributions of parameter estimates and test statistics obtained from MARM. Our new theoretical results show that in MARM, the adaptively weighting idea of the novel PS approach is valid without imposing the propagation condition, which is very difficult to check in real neuroimaging studies. There are several additional theoretical challenges. The first challenge is that the number of voxels, which is much larger than the number of subjects, depends on imaging resolution, not the number of subjects. Thus, we cannot really use the infill asymptotic in the literature of spatial statistics. The second challenge is to choose appropriate kernel functions in constructing adaptive weights, which depend on both the numbers of subjects and voxels.

Section 3.2 of this paper presents MARM just described and establishes the associated theoretical properties. We establish the consistency and asymptotic normality of the adaptive estimator and the asymptotic distribution of the adaptive test statistic for MARM. In Section 3.3, we conduct three sets of simulation studies with the known ground truth to examine the finite sample performance of the adaptive estimate and test statistic in MARM. Section 3.4 illustrates an application of the proposed methods in a real neuroimaging dataset. We present concluding remarks in Section 3.5.

3.2 Multiscale Adaptive Regression Model

3.2.1 Model Formulation

We consider MRI measures in the 3D volume (or on the 2D surface) and clinical variables from n subjects. Without loss of generality, we focus on the 3D volume of MRI measures. For the ith subject, we observe an $mN_D \times 1$ 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 $Y_i(d)$ is an $m \times 1$ vector of MRI measures, $\mathcal{D}$ and d, respectively, represent a 3D volume and a voxel in $\mathcal{D}$ and N_D equals the number of voxels in $\mathcal{D}$. In neuroimaging studies, MRI measures can include the shape representation of the surfaces of cortical or subcortical structures, the determinant of the Jacobian matrices based on the deformation fields estimated by the registration algorithm, functional MRI signals, diffusion tensors, and so on (Ashburner and Friston, 2000; Styner *et al.*, 2005; Thompson and Toga, 2002). Clinical variables often include pedigree information, time, demographic characteristics (e.g., age, gender, height), and diagnostic status, among others.

Statistically, our primary interest is to build the conditional distribution of $\mathbf{Y}_{\mathcal{D}} = \{\mathbf{Y}_{i,\mathcal{D}} : i = 1, \cdots, n\}$ given $\mathbf{X} = \{\mathbf{x}_i : i = 1, \cdots, n\}$, that is, $p(\mathbf{Y}_{\mathcal{D}}|\mathbf{X})$. For a cross-sectional design, it is natural to assume that data from different subjects are independent, that is $p(\mathbf{Y}_{\mathcal{D}}|\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$. Thus, we only need to specify $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$ for each i. However, the number of voxels in each brain region can be more than 500,000 voxels, and at each voxel, the dimension of $Y_i(d)$ can be univariate or multivariate, thus totaling a billion or more data points in an entire study. In addition, imaging data $\mathbf{Y}_{i,\mathcal{D}}$ are spatially correlated in nature, and thus given the large number of voxels on each brain structure, it is statistically challenging to simultaneously model the spatial relationship among all pairs of points.

The voxel-wise approach essentially assumes that

$$p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i) = \prod_{d\in\mathcal{D}} p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d)), \tag{3.1}$$

where $p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d))$ is the marginal density of $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$ and $\boldsymbol{\theta}(d) = (\boldsymbol{\theta}_1(d), \cdots, \boldsymbol{\theta}_p(d))^T$ is a $p \times 1$ vector in an open subset Θ of R^p. Note that due to possible model misspecification, $p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d))$ is only a 'pseudo' density function for $Y_i(d)$. Model (3.1) is general enough to comprise most statistical models in the existing voxel-wise approach. However, since the voxel-wise approach does not account for the spatial nature of neuroimaging data, which often 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 neuroimaging data.

We propose the multiscale adaptive regression model as follows. Assume that for a relatively large radius r_0, $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$ can be well approximated by the product of $p(\{Y_i(d') : \quad d' \in B(d, r_0)\}|\mathbf{x}_i)$, that is

$$p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i) \approx \prod_{d\in\mathcal{D}} p(\{Y_i(d') : \quad d' \in B(d, r_0)\}|\mathbf{x}_i), \tag{3.2}$$

where $B(d, r_0)$ denotes the set of all voxels in a spherical neighborhood of a voxel d with radius r_0. Using all data in all $B(d, r_0)$s can at least preserve the local spatial correlation structure in the imaging data; see panels (a)-(c) in Fig. 3.1. Moreover, since for a given radius r_0, the spherical neighborhoods $B(d, r_0)$ of all voxels are consecutively connected, (3.2) can capture a substantial amount of global spatial information in the neuroimaging data. Statistically, the right hand-side of (3.2) can be regarded as a composite likelihood (Lindsay, 1988; Varin, 2008).

In many neuroimaging studies, our primary interest is to make statistical inference about a vector of parameters of interest, denoted by $\boldsymbol{\theta}(d)$, at each voxel $d \in \mathcal{D}$. It would be very efficient to utilize all data $\{Y_i(d') : \quad d' \in B(d, r_0)\}$ to estimate $\boldsymbol{\theta}(d)$. Instead

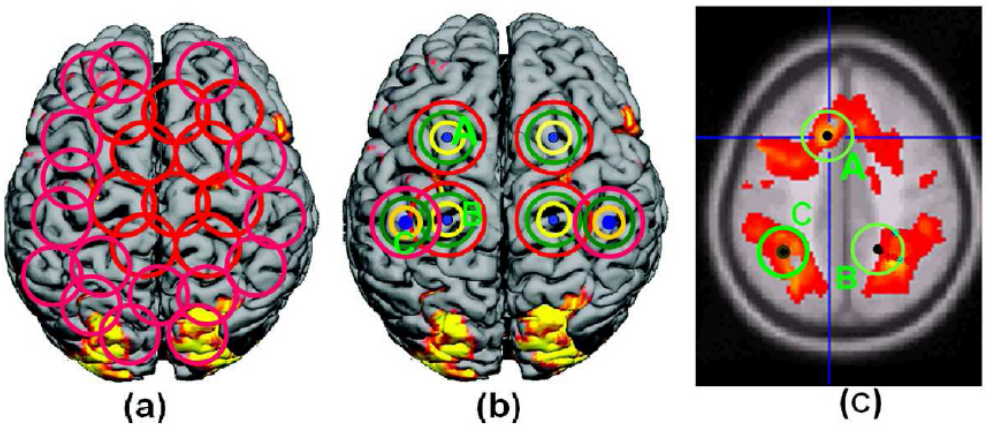

Figure 3.1: Illustrating the key features of the multiscale adaptive regression model. For a relatively large radius r_0, panel (a) shows the overlapping spherical neighborhoods $B(d, r_0)$ of multiple points (or voxels) d on the cortical surface. Panel (b) shows the spherical neighborhoods with four different bandwidths h of the six selected points d on the cortical surface. Panel (c) shows the spherical neighborhoods $B(d, r_0)$ of three selected voxels in a 3D volume, in which voxels A and C are inside the activated regions, whereas voxel B is on the boundary of an activated region.

of specifying spatial correlations among all the $\{Y_i(d') : d' \in B(d, r_0)\}$, assume that $p(\{Y_i(d') : \quad d' \in B(d, r_0)\}|\mathbf{x}_i)$ can be approximated by

$$p(\{\mathbf{Y}_i(d') : \quad d' \in B(d, r_0)\}|\mathbf{x}_i) \approx \prod_{d' \in B(d, r_0)} p(\mathbf{Y}_i(d')|\mathbf{x}_i, \boldsymbol{\theta}(d'))^{\omega(d, d'; r_0)}, \tag{3.3}$$

where $\omega(d, d'; h)$ as a weight function of two voxels and a radius h 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 neuroimaging 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 neuroimaging data are spatially correlated and contain spatially contiguous regions of activation with rather sharp edges. 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. Finally, we obtain an

approximation of $p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i)$ given by

$$p(\mathbf{Y}_{i,\mathcal{D}}|\mathbf{X}_i) \approx \prod_{d\in\mathcal{D}}\{\prod_{d'\in B(d,r_0)} p(\mathbf{Y}_i(d')|\mathbf{x}_i,\boldsymbol{\theta}(d))^{\omega(d,d';r_0)}\}. \tag{3.4}$$

An important issue for MARM is 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 then we adaptively estimate $\boldsymbol{\theta}(d)$ and its associated test statistic. Our multiscale strategy starts with building a sequence of nested spheres with increasing radiuses $h_0 = 0 < h_1 < \cdots < h_S = r_0$ ranging from the smallest scale $h_0 = 0$ to a large scale $h_S = r_0$ at each $d \in \mathcal{D}$ (panel (b) in Fig. 3.1). By setting $\omega(d,d';h_0) = 1$, we can estimate $\boldsymbol{\theta}(d)$ at scale h_0, denoted by $\hat{\boldsymbol{\theta}}(d;h_0)$, and construct a test statistic $W_\mu(d,h_0)$. Then, based on the information contained in $\{\hat{\boldsymbol{\theta}}(d;h_0) : d \in \mathcal{D}\}$, we use some 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 $W_\mu(d,h_s)$ from $h_0 = 0$ to $h_S = r_0$. A path diagram of the multiscale strategy is given below:

$$\begin{array}{ccccccc}
\omega(d,d';h_0) & & \omega(d,d';h_1) & & \cdots & & \omega(d,d';h_S=r_0) \\
\Downarrow & \nearrow & \Downarrow & \nearrow & \cdots & \nearrow & \Downarrow \\
\hat{\boldsymbol{\theta}}(d;h_0) & & \hat{\boldsymbol{\theta}}(d;h_1) & & \cdots & & \hat{\boldsymbol{\theta}}(d;h_S) \\
\Downarrow & & \Downarrow & & \cdots & & \Downarrow \\
W_\mu(d;h_0) & & W_\mu(d;h_1) & & \cdots & & W_\mu(d;h_S)
\end{array} \tag{3.5}$$

At each iteration, the computation involved for 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 MARM sequentially includes more data at each iteration, it will adaptively increase the statistical efficiency in estimating $\boldsymbol{\theta}(d)$ in a homogenous region and decreases the variation of the weights $\omega(d,d';h)$. This multiscale strategy distinguishes MARM from the composite likelihood

methods in the literature (Lindsay, 1988; Varin, 2008).

3.2.2 Estimation and Hypothesis Testing At a Fixed Radius

We present the estimation method and test statistic at each $d \in \mathcal{D}$ for a fixed radius h. Specifically, we consider maximum weighted likelihood estimates of $\boldsymbol{\theta}(d)$ across all voxels $d \in \mathcal{D}$ given the current weights $\{\omega(d, d'; h) : d, d' \in \mathcal{D}\}$. For the sphere with the radius h of the voxel d, the weighted quasi-likelihood function $\ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}})$ is given by

$$\ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}}) = \sum_{i=1}^{n} \sum_{d' \in B(d,h)} \tilde{\omega}(d, d'; h) \log p(Y_i(d')|\mathbf{x}_i, \boldsymbol{\theta}(d)), \tag{3.6}$$

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

$$\hat{\boldsymbol{\theta}}(d, h) = \text{argmax}_{\boldsymbol{\theta}(d)} n^{-1} \ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}}). \tag{3.7}$$

Numerically, we use various algorithms, such as Newton-type algorithms, to estimate $\hat{\boldsymbol{\theta}}(d, h)$. Throughout the paper, the Newton-Raphson algorithm is used to calculate $\hat{\boldsymbol{\theta}}(d, h)$ by iterating

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

where $\partial_{\boldsymbol{\theta}(d)}$ and $\partial^2_{\boldsymbol{\theta}(d)}$ denote, respectively, the first- and second-order partial derivatives with respect to $\boldsymbol{\theta}(d)$ evaluated at $\hat{\boldsymbol{\theta}}(d, h)^{(t)}$. To stabilize the Newton-Raphson algorithm, we approximate $-\partial^2_{\boldsymbol{\theta}(d)} \ell_n(\hat{\boldsymbol{\theta}}(d, h)^{(t)}; h, \tilde{\boldsymbol{\omega}})$ by $E[-\partial^2_{\boldsymbol{\theta}(d)} \ell_n(\hat{\boldsymbol{\theta}}(d, h)^{(t)}; h, \tilde{\boldsymbol{\omega}})]$. We stop the Newton-Raphson algorithm 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, $\text{Cov}[\hat{\boldsymbol{\theta}}(d, h)]$ can

be approximated by

$$\text{Cov}[\hat{\boldsymbol{\theta}}(d,h)] \approx \Sigma_n(\hat{\boldsymbol{\theta}}(d,h)) = [\Sigma_{n,1}(\hat{\boldsymbol{\theta}}(d,h))]^{-1}\Sigma_{n,2}(\hat{\boldsymbol{\theta}}(d,h))[\Sigma_{n,1}(\hat{\boldsymbol{\theta}}(d,h))]^{-1}, \quad (3.8)$$

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

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

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 (or function) across diagnostic groups or the detection of a change in brain structure (or function) across time (Styner *et al.*, 2005; Thompson and Toga, 2002; Zhu *et al.*, 2007a). These questions of interest usually can be formulated as testing hypotheses about $\boldsymbol{\theta}(d)$ as follows:

$$H_{0,\mu} : R(\boldsymbol{\theta}(d)) = \mathbf{b}_0 \quad \text{vs.} \quad H_{1,\mu} : R(\boldsymbol{\theta}(d)) \neq \mathbf{b}_0, \quad (3.9)$$

where $\mu = R(\boldsymbol{\theta}(d))$ is an $r \times 1$ vector function of the $k-$vector $\boldsymbol{\theta}(d)$ with $r \geq k$ 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 Wald test statistic $W_\mu(d,h)$, which is given by

$$[R(\hat{\boldsymbol{\theta}}(d;h)) - \mathbf{b}_0]^T\{\partial_{\boldsymbol{\theta}(d)}R(\hat{\boldsymbol{\theta}}(d;h))\hat{\Sigma}_n(\hat{\boldsymbol{\theta}}(d;h))\partial_{\boldsymbol{\theta}(d)}R(\hat{\boldsymbol{\theta}}(d;h))^T\}^{-1}[R(\hat{\boldsymbol{\theta}}(d;h)) - \mathbf{b}_0]. \quad (3.10)$$

To test whether $H_{0,\mu}$ holds in all voxels of the region under study, we may consider various statistical methods including the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) and the random field theory (Worsley *et al.*, 2004). In most applications, we are interested in testing $R(\boldsymbol{\theta}(d)) = \mathbf{R}_0\boldsymbol{\theta}(d)$ for a given $r \times k$ matrix $\mathbf{R}_0$. For simplicity, we only consider testing $H_0 : \mathbf{R}_0\boldsymbol{\theta}(d) = \mathbf{b}_0$ from now on.

3.2.3 Adaptive Estimation and Testing Procedure

We develop an adaptive estimation and testing (AET) procedure evolving from the smallest scale $h_0 = 0$ to the largest scale $h_S = r_0$ for MARM. The AET procedure starts with individual voxel $d \in \mathcal{D}$ and then successively increases the radius (or bandwidth) h_s of a spherical neighborhood around each $d \in \mathcal{D}$. For a given $d \in \mathcal{D}$, 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 utilize all data in $B(d, h_s)$ and the adaptive weights in all these voxels to obtain updated estimates $\hat{\boldsymbol{\theta}}(d, h_s)$ and $W_\mu(d, h_s)$ at each voxel $d \in \mathcal{D}$, respectively.

The AET procedure consists of four key steps: initialization, weights adaptation, estimation, and stopping. In the initialization step (i), we prefix a geometric series $\{h_s = c_h^s : s = 1, \cdots, S\}$ of radiuses with $h_0 = 0$, where $c_h \in (1, 2)$, say $c_h = 1.25$. At each voxel d, we calculate the MWQL estimate $\hat{\boldsymbol{\theta}}(d, h_0)$ and its associated Wald test statistic $W_\mu(d, h_0)$, which are the same as those from the voxel-wise approach. We then set $s = 1$ and $h_1 = c_h$.

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

$$\omega(d, d'; h_s) = K_{loc}(||d - d'||_2/h_s) K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_{s-1})/C_n), \tag{3.11}$$

where $K_{loc}(u)$ and $K_{st}(u)$ are two kernel functions with compact support such that both $K_{loc}(u)$ and $K_{st}(u)$ decrease to zero as u increases, C_n is a number, which may be associated with n, and $||\cdot||_2$ denotes the Euclidean norm of a vector (or a matrix). Moreover, $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$ denotes a weighted function of the MWQL estimates of $\{\boldsymbol{\theta}(d) : d \in \mathcal{D}\}$ calculated as the radius equals h_{s-1}. The adaptive weight $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_{s-1})/C_n)$ down-weight the role of a voxel $d' \in B(d, h_s)$ in $\ell_n(\boldsymbol{\theta}(d); h_s, \tilde{\boldsymbol{\omega}})$ if the value of $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$ is large. The weights $K_{loc}(||d - d'||_2/h_s)$ give less weight to the voxel $d' \in B(d, h_s)$, whose location is far from the voxel d.

In the estimation step (iii), for the radius h_s, we calculate $\hat{\boldsymbol{\theta}}(d, h_s)$ and $W_\mu(d, h_s)$, which are defined in (3.7) and (3.10), respectively, at each voxel $d \in \mathcal{D}$.

In the stopping step (iv), when $s = S$, we compute the p-values for $W_\mu(d, h_S)$, apply either FDR or RFT to detect significant voxels and then stop. Otherwise, we 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 5, such that the largest spherical neighborhood of each voxel only contains a relatively small number of voxels compared with the whole volume.

Remark 1. We have developed the AET procedure for spatially and adaptively carrying out statistical inference on all components of $\boldsymbol{\theta}(d)$ in the 3D volume (or 2D surface) as a whole. However, in many applications, $\boldsymbol{\theta}(d)$ may be decomposed as $(\boldsymbol{\theta}_1(d)^T, \boldsymbol{\theta}_2(d)^T)^T$, in which $\boldsymbol{\theta}_1(d)$ is the parameter of interest and $\boldsymbol{\theta}_2(d)$ is the nuisance parameter. We can also develop the AET procedure for $\boldsymbol{\theta}_1(d)$ only. Specifically, we can calculate $\hat{\boldsymbol{\theta}}(d, h_0)$ and then fix $\boldsymbol{\theta}_2(d)$ at $\hat{\boldsymbol{\theta}}_2(d, h_0)$ after the initialization step (i). In this way, we only update $\boldsymbol{\theta}_1(d)$ and calculate adaptive weights based on estimates of $\boldsymbol{\theta}_1(d)$ at each iteration.

Remark 2. Setting $\hat{\boldsymbol{\theta}}(d, h_s)^{(0)} = \hat{\boldsymbol{\theta}}(d, h_{s-1})$ for each $s > 0$ is an efficient way of selecting the initial value $\hat{\boldsymbol{\theta}}(d, h_s)^{(0)}$ in the Newton-Raphson algorithm. Since the AET procedure always downweights voxel $d' \in B(d, h)$ in $\ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}})$ when the value of $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$ 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, and thus the additional computational time for MARM is very light compared to the voxel-wise approach.

Remark 3. There are two different kernel functions in the AET procedure. The $K_{loc}(u)$ is a regular kernel function for smoothing the smoothed curves or surfaces. Some common choices of $K_{loc}(u)$ include the Gaussian kernel and the Epanechnikov kernel (Tabelow *et al.*, 2006; Polzehl and Spokoiny, 2000, 2006). Without loss of generality, we use $K_{loc}(u) = (1 - u)_+$. The $K_{st}(u)$ is the kernel function for downweighting the voxels

that are dissimilar to voxel d during the process of making inference on $\boldsymbol{\theta}(d)$. In practice, we set $K_{st}(u) = e^{-u}\mathbf{1}(u \leq 5)$ and

$$D_{\boldsymbol{\theta}}(d, d'; h_{s-1}) = [\hat{\boldsymbol{\theta}}(d, h_{s-1}) - \hat{\boldsymbol{\theta}}(d', h_{s-1})]^T \hat{\Sigma}(\hat{\boldsymbol{\theta}}(d; h_{s-1}))^{-1} [\hat{\boldsymbol{\theta}}(d, h_{s-1}) - \hat{\boldsymbol{\theta}}(d', h_{s-1})]. \tag{3.12}$$

The $D_{\boldsymbol{\theta}}(d, d'; h_{s-1})$ in (3.12) is close to $s_{ij}^{(k)}$ in equation (14) of Tabelow *et al.* (2008). Moreover, although different choices of $K_{loc}(u)$ and $K_{st}(u)$ have been suggested in the original PS approach (Polzehl and Spokoiny 2000, 2006), it is unclear what kinds of kernel functions should be chosen in MARM both theoretically and numerically.

Remark 4. A crucial issue in the MARM approach is to appropriately select C_n in (3.11). In the PS procedure, various choices of C_n including the logarithm of the number of voxels in $B(d, h)$ and the quantile of the χ^2 distribution have been suggested (Polzehl and Spokoiny 2000, 2006; Katkovnik and Spokoiny 2008). We will formally examine what kinds of C_n should be used in MARM.

3.2.4 Theoretical Properties

Throughout the paper, we only consider the asymptotic properties of $\hat{\boldsymbol{\theta}}(d, h_s)$ and $W_\mu(d, h_s)$ for finite number of iterations and bounded r_0 for MARM. We assume that the number of voxels in brain volume does not increase with the sample size, since the resolution of a given imaging dataset is always fixed. One might attempt to consider 'infill asymptotics' in spatial statistics, in which the brain volume is fixed and the voxel size approaches zero as the sample size increases, but the voxel size in neuroimaging is associated with imaging resolution, not the sample size. Therefore, we will not consider the infill asymptotics.

We establish consistency and asymptotically normality of $\hat{\boldsymbol{\theta}}(d, h)$ and $W_\mu(d, h)$ for each h obtained from the AET procedure in Section 3.2.4. We first discuss the case with fixed weights $\omega(d, d'; h)$ for a fixed scale h. Let $Y_i(d, h) = (Y_i(d') : d' \in B(d, h))$ for $i =$

$1, \cdots, n$. Without loss of generality, we assume that the $(Y_i(d,h), \mathbf{x}_i)$ are independently and identically distributed as the true density $p(Y(d,h), \mathbf{x})$. According to (3.6), the MWQL estimator $\hat{\boldsymbol{\theta}}(d,h)$ maximizes the function $n^{-1}\ell_n(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}})$, which converges to

$$M(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}}) = \sum_{d' \in B(d,h)} \tilde{\omega}(d, d'; h) E[\log p(Y(d')|\mathbf{x}, \boldsymbol{\theta}(d))] \tag{3.13}$$

in probability (or almost surely) under some mild conditions as $n \to \infty$, where the expectation is taken with respect to $p(Y(d,h), \mathbf{x})$. Under some identifiability conditions, $\hat{\boldsymbol{\theta}}(d;h)$ converges to $\boldsymbol{\theta}_*(d;h)$, which maximizes $M(\boldsymbol{\theta}(d); h, \tilde{\boldsymbol{\omega}})$ (van der Vaart, 1998). When $h = 0$, $\boldsymbol{\theta}_*(d;0) = \boldsymbol{\theta}_*(d)$ is the 'pseudo' true value in voxel d and the parameter of interest. When $h > 0$, $\boldsymbol{\theta}_*(d;h)$ can only be regarded as a weighted combination of all $\boldsymbol{\theta}_*(d')$ for $d' \in B(d,h)$. In a homogeneous region, that is $\boldsymbol{\theta}_*(d') = \boldsymbol{\theta}_*(d)$, $\boldsymbol{\theta}_*(d;h) = \boldsymbol{\theta}_*(d)$ even for $h > 0$. However, in a nonhomogeneous region, an arbitrary set of weights $\omega(d, d'; h)$ can lead to undesirable consequences, such as smoothing out the boundary of activated regions and reducing statistical power in detecting activated regions.

We need to address several important questions for MARM with stochastic adaptive weights. A critical question is that what kinds of stochastic weights can automatically incorporate the 'good' information and prevent the 'bad' information from neighboring voxels. By appropriately utilizing information from neighboring voxels, the AET procedure can dramatically increase the accuracy and efficiency in estimating $\boldsymbol{\theta}_*(d)$ in each voxel. Another important question is whether the stochastic weights chosen can ensure the consistency and asymptotical normality of $\hat{\boldsymbol{\theta}}(d,h)$ at each fixed scale h. To have a better understanding of the AET procedure, we focus on the asymptotic behavior of the adaptive weights as $s = 1$ and then we discuss the scenario with $s > 1$.

We obtain the following theorems, whose detailed assumptions and proofs can be found in the Appendix.

Theorem 1. If assumptions (C1)-(C7) in the Appendix are true, then we have

(a) $\hat{\boldsymbol{\theta}}(d, h_0)$ converges to $\boldsymbol{\theta}_*(d)$ in probability;

(b) $\{\Sigma_{n,2}(\hat{\boldsymbol{\theta}}(d,h_0))\}^{-1/2}\Sigma_{n,1}(\hat{\boldsymbol{\theta}}(d,h_0))[\hat{\boldsymbol{\theta}}(d,h_0)-\boldsymbol{\theta}_*(d)] \to^L N(0,\mathbf{I}_p)$*;*

(c) $D_{\boldsymbol{\theta}}(d,d';h_0)$ *and* $K_{st}(D_{\boldsymbol{\theta}}(d,d';h_0)C_n^{-1})$ *can be, respectively, approximated by*

$$\begin{aligned} D_{\boldsymbol{\theta}}(d,d';h_0) &= \mathbf{1}(\triangle_*(d,d')=\mathbf{0})\times O_p(\log(N(\mathcal{D}))) + \mathbf{1}(\triangle_*(d,d')\neq\mathbf{0}) \\ &\quad n||[\Sigma_*(d,h)]^{-1/2}[\triangle_*(d,d')+O_p(\sqrt{\log(N(\mathcal{D}))/n})]||_2^2, \end{aligned} \tag{3.14}$$

$$\begin{aligned} K_{st}(D_{\boldsymbol{\theta}}(d,d';h_0)C_n^{-1}) &= \mathbf{1}(\triangle_*(d,d')\neq\mathbf{0})K_{st}(C_n^{-1}nO_p(1)) \\ &+ \mathbf{1}(\triangle_*(d,d')=\mathbf{0})K_{st}(\log(N(\mathcal{D}))C_n^{-1}O_p(1)), \end{aligned} \tag{3.15}$$

where $\triangle_*(d,d')=\boldsymbol{\theta}_*(d)-\boldsymbol{\theta}_*(d')$ *and* $N(\mathcal{D})$ *denotes the number of voxels in* $\mathcal{D}$*;*

(d) For any $\epsilon_0>0$*,*

$$\lim_{n\to\infty} P(|K_{st}(D_{\boldsymbol{\theta}}(d,d';h_0)/C_n)-\mathbf{1}(\triangle_*(d,d')=\mathbf{0})|>\epsilon_0)=0.$$

Theorem 1 (a) and (b) characterize the asymptotic behavior of $D_{\boldsymbol{\theta}}(d,d';h_0)$ and $K_{st}(D_{\boldsymbol{\theta}}(d,d';h_0)/C_n)$. Theorem 1 (c) and (d) show that if the two voxels d and d' have the same true values, then $K_{st}(D_{\boldsymbol{\theta}}(d,d';h_0)/C_n)$ and $\omega(d,d';h_0)$ converges to 1 and $K_{loc}(||d-d'||_2/h_1)$, respectively. However, if the two voxels d and d' substantially differ from each other, then $K_{st}(D_{\boldsymbol{\theta}}(d,d';h_0)/C_n)$ imposes an decreasing weight on the voxel d'. Thus, $K_{st}(D_{\boldsymbol{\theta}}(d,d';h_0)/C_n)$ can efficiently incorporate information from 'good' voxels, whereas it prevents incorporating information from 'bad' voxels.

For $h>0$, we can also establish important theoretical results to characterize the nice behavior 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 Appendix are true. We have the following results for the MARM:

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

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

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

Theorem 2 shows that the AET procedure has several remarkable features. Theorem 2 (a) ensures that $\hat{\boldsymbol{\theta}}(d,h)$ is a consistent estimate of $\boldsymbol{\theta}_*(d)$ for the adaptive weights in (3.10) for any h. Theorem 2 (b) ensures that $\hat{\boldsymbol{\theta}}(d,h)$ is a $\sqrt{n}$ estimate of $\boldsymbol{\theta}_*(d)$. Theorem 3 (c) ensures that the Wald test statistic $W_\mu(d,h_s)$ is asymptotically $\chi^2(r)$ distributed under the null hypothesis $\mathbf{R}_0\boldsymbol{\theta}_*(d)=\mathbf{b}_0$. However, for small sample size 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 $\frac{r(n-1)}{n-r}F^{1-\alpha}_{r,n-r}$, where $F^{1-\alpha}_{r,n-r}$ is the upper α-percentile of the $F_{r,n-r}$ distribution. That is, we reject H_0 if $W_\mu(d,h)\geq\frac{r(n-1)}{n-r}F^{1-\alpha}_{r,n-r}$, and do not reject H_0 otherwise.

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

Corollary 1. Suppose assumptions (C1)-(C6) in the Appendix are true and $C_n=O(1)$ *and* $\lim_{n\to\infty}\log(N(\mathcal{D}))/n=0$. *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* $\mathbf{R}_0\boldsymbol{\theta}_*(d)=\mathbf{b}_0$ *is true.*

Corollary 1 shows that bounded C_n can lead to several unpleasant consequences. Although bounded C_n has been proposed in the PS approach to smooth the parameter estimates from linear models, it is the first time that we establish the consistency of $\hat{\boldsymbol{\theta}}(d,h)$ as an estimate of $\boldsymbol{\theta}_*(d)$ under a general setup. However, if there is a voxel $d'\in B(d,h)/\{d\}$ such that $\boldsymbol{\theta}_*(d)=\boldsymbol{\theta}_*(d')$, Corollary 1 (b) shows that $\hat{\boldsymbol{\theta}}(d,h)$ is not asymptotically normal and the Wald test statistic $W_\mu(d,h_s)$ is not asymptotically $\chi^2(r)$ distributed under the null hypothesis $R\boldsymbol{\theta}_*(d)=\mathbf{b}_0$. Thus, we cannot directly calibrate $W_\mu(d,h)$ using the critical values of $\chi^2(r)$.

3.2.5 Multiscale Adaptive Generalized Linear Models

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

$$\exp\left\{\tau(d)[Y_i(d)\eta_i(\boldsymbol{\beta}(d)) - b(\eta_i(\boldsymbol{\beta}(d)))] + c(Y_i(d), \tau(d))\right\}, \tag{3.16}$$

$i = 1, \ldots, n$, indexed by the canonical parameter η_i and the scale parameter $\tau(d)$, where the functions $b(\cdot)$ and $c(\cdot, \cdot)$ determine a particular distributional family in the class, such as the binomial, normal or Poisson distributions. Furthermore, the η_i's satisfy the equations $\eta_i(\boldsymbol{\beta}(d)) = \eta(\mu_i(d))$, $i = 1, \ldots, n$, and $\mu_i(d) = g(\mathbf{x}_i^T\boldsymbol{\beta}(d))$, where $\boldsymbol{\theta}(d) = (\boldsymbol{\beta}(d), \tau(d))$ and $g(\cdot)$ is an known and monotonic link function and $\boldsymbol{\beta}(d)$ is a $p-1$-dimensional vector of regression coefficients. The GLMs include many well-known regression models, such as normal linear regression, logistic and probit regression, Poisson regression, gamma regression, and some proportional hazards models (McCullagh and Nelder, 1989).

We develop the multiscale GLM and present several key formula below. Our primary interest is $\boldsymbol{\beta}(d)$, so $\tau(d)$ is fixed at $\hat{\tau}(d, h_0)$ from now on. For the scale h, we define

$$\begin{aligned}
&\tilde{\omega}(d, d'; h) = \tau(d')\omega(d, d'; h)/\overline{\omega}(d; \boldsymbol{\omega}, h),\\
&c_i(\tau(d); \tilde{\boldsymbol{\omega}}, h) = \sum_{d' \in B(d,h)} \tilde{\omega}(d, d'; h) c(Y_i(d'), \tau(d')),\\
&\overline{\omega}(d; \boldsymbol{\omega}, h) = \sum_{d' \in B(d,h)} \tau(d')\omega(d, d'; h), \quad \text{and} \quad Y_i(d; \tilde{\boldsymbol{\omega}}, h) = \sum_{d' \in B(d,h)} \tilde{\omega}(d, d'; h) Y_i(d'),
\end{aligned}$$

in which $Y_i(d; \tilde{\boldsymbol{\omega}}, h)$ is a weighted response at voxel d for $i = 1, \cdots, n$ at the scale h. The weighted quasi-likelihood function at voxel d for the scale h is given by

$$\sum_{i=1}^{n} \{Y_i(d; \tilde{\boldsymbol{\omega}}, h)\eta_i(\boldsymbol{\beta}(d)) - b(\eta_i(\boldsymbol{\beta}(d)))\} + \sum_{i=1}^{n} c_i(\tau(d); \tilde{\boldsymbol{\omega}}, h). \tag{3.17}$$

With some calculations, we get

$$\begin{aligned}
\partial_{\boldsymbol{\beta}(d)}\ell_n(\boldsymbol{\theta}(d);h,\tilde{\boldsymbol{\omega}}) &= \sum_{i=1}^{n}\{Y_i(d;\boldsymbol{\omega},h)-\dot{b}(\eta_i(\boldsymbol{\beta}(d)))\}\partial_{\boldsymbol{\beta}(d)}\eta_i(\boldsymbol{\beta}(d)),\\
-\partial^2_{\boldsymbol{\beta}(d)}\ell_n(\boldsymbol{\theta}(d);h,\tilde{\boldsymbol{\omega}}) &= \sum_{i=1}^{n}\ddot{b}(\eta_i(\boldsymbol{\beta}(d)))[\partial_{\boldsymbol{\beta}(d)}\eta_i(\boldsymbol{\beta}(d))]^{\otimes 2}\\
&- \sum_{i=1}^{n}\{Y_i(d;\tilde{\boldsymbol{\omega}},h)-\dot{b}(\eta_i(\boldsymbol{\beta}(d)))\}\partial^2_{\boldsymbol{\beta}(d)}\eta_i(\boldsymbol{\beta}(d)),
\end{aligned}$$

where $\dot{b}(t)=db(t)/dt$ and $\ddot{b}(t)=d^2b(t)/dt^2$. Based on these preparations, we can develop the Newton-Raphson algorithm and calculate $\Sigma_{n,1}(\boldsymbol{\beta}(d))$ and $\Sigma_{n,2}(\boldsymbol{\beta}(d))$, which lead an approximation of $\mathrm{Cov}(\hat{\boldsymbol{\beta}}(d,h))$. For the canonical link $\eta_i(\boldsymbol{\beta}(d))=\mathbf{x}_i^T\boldsymbol{\beta}(d)$, all formula such as $\partial_{\boldsymbol{\beta}(d)}\eta_i(\boldsymbol{\beta}(d))=\mathbf{x}_i$ can be further simplified.

As an illustration, we focus on the well-known linear model and develop a multiscale linear model. In particular, we examine the asymptotic properties of $\hat{\boldsymbol{\beta}}(d,h)$. Assume that $Y_i(d)=\mathbf{x}_i^T\boldsymbol{\beta}(d)+\epsilon_i(d)$, where $\epsilon_i(d)\sim N(0,\tau(d)^{-1})$. With simple calculation, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(d,h) &= (\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2})^{-1}\sum_{i=1}^{n}\mathbf{x}_iY_i(d;\tilde{\boldsymbol{\omega}},h)\\
&= \sum_{d'\in B(d,h)}\tilde{\omega}(d,d';h)\boldsymbol{\beta}_*(d')+\sum_{d'\in B(d,h)}\tilde{\omega}(d,d';h)(\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2})^{-1}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i(d'), \qquad (3.18)\\
\mathrm{Cov}[\hat{\boldsymbol{\beta}}(d,h)] &\approx (\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2})^{-1}\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2}\hat{\epsilon}_i(d;\omega,h)^2(\sum_{i=1}^{n}\mathbf{x}_i^{\otimes 2})^{-1}, \qquad (3.19)
\end{aligned}$$

where $\hat{\epsilon}_i(d;\boldsymbol{\omega},h)=\sum_{d'\in B(d,h)}\tilde{\omega}(d,d';h)[Y_i(d')-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}(d',h)]$. Although Tabelow *et al.* (2006) have obtained the same $\hat{\boldsymbol{\beta}}(d,h)$ as in (3.18), the MARM developed here has several advantages over the PS approach. We will show below that $\hat{\boldsymbol{\beta}}(d,h)$ based on the adaptive weights in the PS approach may not be asymptotically normal. The covariance estimate of $\hat{\boldsymbol{\beta}}(d,h)$ in (3.19) has a simpler form compared to that in Tabelow *et al.* (2006), in which they first estimated a spatial correlation factor and applied it to all

voxels. In real neuroimaging studies, it is unrealistic to assume the homogeneous spatial correlation, while it is unnecessary to estimate the spatial correlation. We obtain the following results for linear model. For simplicity, we assume that all $\tau(d)$ are known.

Theorem 3. (a) If Assumptions (C1), (C2), (C6) and (C7) are true and $E[||\mathbf{x}||_2^2] < \infty$ *and* $E[\max_{d\in\mathcal{D}} |\epsilon(d)| \times ||\mathbf{x}||_2^2] < \infty$*, then* $\sqrt{n}[\hat{\boldsymbol{\beta}}(d,h) - \boldsymbol{\beta}_*]$ *is asymptotically equivalent to*

$$A_1(d;h) = \frac{\sum_{d'\in B(d,h)} C(d,d';h)\tau(d')E[\mathbf{x}^{\otimes 2}]^{-1}n^{-1/2}\sum_{i=1}^n \mathbf{x}_i\epsilon_i(d')}{\sum_{d'\in B(d,h)} C(d,d';h)\tau(d')}, \tag{3.20}$$

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

$$\frac{\sum_{d'\in B(d,h)} C(d,d';h)\tau(d')E[\mathbf{x}^{\otimes 2}]^{-1/2}Z(d')}{\sum_{d'\in B(d,h)} C(d,d';h)\tau(d')}, \tag{3.21}$$

where $(Z(d) : d \in B(d,h))$ *is a Gaussian vector with mean zero and a covariance structure* $Cov(Z(d)) = \tau(d)^{-1}\mathbf{I}_{p-1}$ *and* $Cov(Z(d), Z(d')) = E[\epsilon_1(d)\epsilon_1(d')]\mathbf{I}_{p-1}$.

(b) If Assumptions (C1), (C2) and (C6) are true and $C_n = O(1)$ *and* $\lim_{n\infty} \log(N(\mathcal{D}))/n = 0$*, then* $\sqrt{n}[\hat{\boldsymbol{\beta}}(d,h_1) - \boldsymbol{\beta}_*]$ *is asymptotically equivalent to*

$$A_2(d;h_1) = \frac{\sum_{d'\in B(d,h_1)} C(d,d';h_1)K_{st}(\mathcal{E}_n(d,d'))\tau(d')E[\mathbf{x}^{\otimes 2}]^{-1}n^{-1/2}\sum_{i=1}^n \mathbf{x}_i\epsilon_i(d')}{\sum_{d'\in B(d,h_1)} C(d,d';h_1)K_{st}(\mathcal{E}_n(d,d'))\tau(d')}, \tag{3.22}$$

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

$$\frac{\sum_{d'\in B(d,h)} C(d,d';h_1)K_{st}(\tau(d)tr\{[Z(d)-Z(d')]^{\otimes 2}\})\tau(d')E[\mathbf{x}^{\otimes 2}]^{-1/2}Z(d')}{\sum_{d'\in B(d,h)} C(d,d';h_1)K_{st}(\tau(d)tr\{[Z(d)-Z(d')]^{\otimes 2})\}\tau(d')}. \tag{3.23}$$

Theorem 3 first time gives a theoretical justification of the multiscale adaptive linear model in Tabelow *et al.* (2006). Theorem 3 (a) and (b) formally characterize the key differences between bounded and unbounded C_n in general linear model. Theorem 3 (a) shows that for certain unbounded C_n, the asymptotic distributions of $\hat{\boldsymbol{\beta}}(d,h)$ are always normally distributed. For bounded C_n, however, Theorem 3 (b) only gives the

asymptotic distribution of $\hat{\boldsymbol{\beta}}(d, h_1)$, which may not be normally distributed when there is a voxel $d' \in B(d, h_1)$ being close to the voxel d. Particularly, the covariance estimate $\hat{\boldsymbol{\beta}}(d, h)$ in Tabelow *et al.* (2006) may not be valid for bounded C_n even for $h = h_1$.

3.3 Simulation Studies

We conducted three sets of Monte Carlo simulations to examine the finite sample performance of $\hat{\boldsymbol{\beta}}(d, h)$ and $W_\mu(d, h)$ with respect to different scales h at the levels of a single voxel and an entire brain region. The first two were based on simulated data on the 2D surface with the known ground truth. The third one was based on simulated MRI datasets in the 3D volume with the known ground truth.

3.3.1 Simulation Studies Part I

We simulated data at all $m = 4002$ points on the surface of a hippocampus for n subjects. At a given voxel d in $\mathcal{D}$, $y_i(d)$ was simulated according to $y_i(d) = \mathbf{x}_i^T \boldsymbol{\beta}(d) + \epsilon_i(d)$ for $i = 1, \cdots, n$, where $\boldsymbol{\beta}(d) = (\beta_1(d), \beta_2(d), \beta_3(d))^T$, $\mathbf{x}_i = (1, x_{i2}, x_{i3})^T$ and the $\epsilon_i(d)$ were independently generated from $N(0, 1)$. We set $n = 60$ and $n = 80$. We generated x_{i2} independently from a Bernoulli distribution with the probability of success being 0.5, and generated x_{i3} independently from the uniform distribution in [0, 1]. The x_{i2} and x_{i3} were chosen to represent group identity and standardized age, respectively. We also created ROI1 and ROI2, which are two nested circles with radius at 3 and 5, respectively, and labeled the region outside of ROI1 and ROI2 as ROI3. We set $\beta_2(d)$ as 0 in ROI3, 1 in ROI2, and 2 in ROI3, respectively (Fig. 3.2(a)).

We fitted the linear model $y_i(d) = \mathbf{x}_i^T \boldsymbol{\beta}(d) + \epsilon(d)$, where $\epsilon_i(d) \sim N(0, \sigma^2(d))$. The $\boldsymbol{\theta}(d)$ includes $\boldsymbol{\beta}(d)$ and $\sigma^2(d)$. We used MARM to calculate adaptive parameter estimates across all voxels at 6 different scales. Since our primary interest is $\boldsymbol{\beta}(d)$ and $\sigma^2(d)$ was treated as nuisance parameters and fixed at $\hat{\sigma}^2(d, h_0)$ after the h_0-th iteration. In each

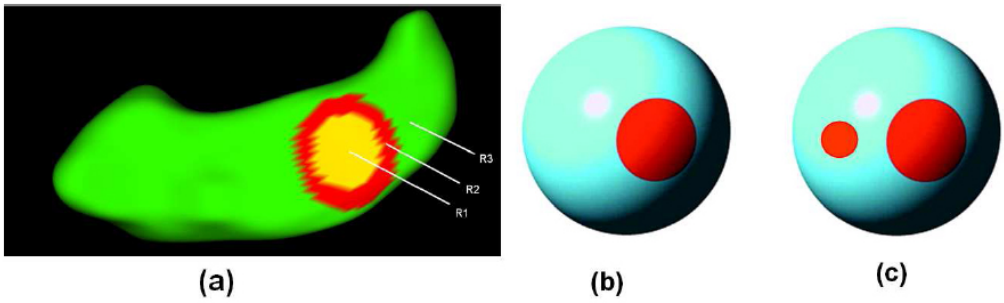

Figure 3.2: Setups for simulation studies parts I and II: (a) three regions of interest ($R1$: ROI1 with yellow color; $R2$: ROI2 with red color; $R3$: ROI3 with green color) on a reference hippocampus; (b) a reference sphere with a red ROI; (c) a reference sphere with two red ROIs.

Table 3.1: Bias ($\times 10^{-2}$), RMS($\times 10^{-2}$), SD ($\times 10^{-2}$), and RS of $\boldsymbol{\beta}$ parameters. BIAS denotes the bias of the mean of the MARM estimates; RMS denotes the root-mean-square error; SD denotes the mean of the standard deviation estimates; RS denotes the ratio of RMS over SD. sample size=60.

	β_1				β_2				β_3			
	BIAS	RMS	SD	RS	BIAS	RMS	SD	RS	BIAS	RMS	SD	RS
						$n = 60$						
ROI1: h_0	4.1	1.61	1.88	0.86	-2.1	1.35	1.19	1.12	-2.1	1.35	1.19	1.13
ROI1: h_5	4.3	0.62	0.59	1.05	-2.7	0.41	0.41	0.99	-5.7	0.79	0.85	0.92
ROI2: h_0	0.66	1.97	1.86	1.06	0.69	1.16	1.18	0.98	3.3	2.71	2.54	1.06
ROI2: h_5	0.41	0.68	0.60	1.14	0.61	0.44	0.38	1.14	3.1	0.81	0.82	0.99
ROI3: h_0	0.27	1.92	1.85	1.04	-0.19	1.28	1.20	1.06	-0.58	2.61	2.53	1.03
ROI3: h_5	0.30	0.55	0.51	1.09	-0.13	0.36	0.33	1.06	-0.65	0.74	0.69	1.07
						$n = 80$						
ROI1: h_0	-2.4	1.56	1.589	0.98	-0.56	1.05	1.06	0.99	1.8	2.37	2.16	1.10
ROI1: h_5	-3.1	0.60	0.53	1.14	-0.83	0.35	0.35	0.99	2.5	0.71	0.71	1.00
ROI2: h_0	-0.66	1.83	1.64	1.11	1.9	1.28	1.05	1.23	0.99	2.36	2.20	1.08
ROI2: h_5	-0.70	0.56	0.54	1.04	1.8	0.43	0.34	1.26	1.0	0.64	0.71.	0.89
ROI3: h_0	-0.12	1.69	1.66	1.02	-0.12	1.69	1.66	1.02	0.17	2.29	2.21	1.03
ROI3: h_5	-0.05	0.48	0.46	1.05	-0.04	0.32	0.29	1.09	0.11	0.66	0.61	1.08

ROI, we calculated the bias, the empirical standard errors (RMS), and the mean of the standard error estimates (SD) based on the results from the 100 simulated hippocampus data sets. We observed the following results. The biases are similar at h_0 and h_5. The RMS and SD at h_5 are much smaller than those at h_0. In addition, the RMS and its corresponding SD are relatively close to each other at both h_0 and h_5 scales in each of the three ROIs (Table 3.1). As expected, increasing n decreases the bias, RMS and SD of parameter estimates.

3.3.2 Simulation Studies Part II

In this simulation, we simulated data at all $m = 2064$ points on the surface of a reference sphere for n subjects. At a given voxel d in $\mathcal{D}$, a 2×1 vector $y_i(d)$ was simulated according to $y_i(d) = \boldsymbol{B}(d)\mathbf{x}_i + \epsilon_i(d)$, where $\boldsymbol{B}(d) = (\beta_{jk}(d))$ is a 2×3 matrix, $\mathbf{x}_i = (1, x_{i2}, x_{i3})^T$ and the $\epsilon_i(d)$ were independently generated from $N(\mathbf{0}, \mathbf{I}_2)$, in which $\mathbf{I}_2$ is a 2×2 identity matrix. We generated x_{i2} independently from a Bernoulli distribution with an equal probability and generated x_{i3} independently from the uniform distribution in [0, 1]. We set $n = 20$, $n = 40$ and $n = 60$.

We fitted the linear model $y_i(d) = \boldsymbol{B}(d)\mathbf{x}_i + \epsilon_i(d)$, where $\epsilon_i(d) \sim N(0, \Sigma_e(d))$. The $\boldsymbol{\theta}(d)$ includes $\boldsymbol{B}(d)$ and the elements in $\Sigma_e(d)$. Since our primary interest is $\boldsymbol{B}(d)$ and the elements in $\Sigma_e(d)$ was treated as nuisance parameters, we fixed $\Sigma_e(d) = \hat{\Sigma}_e(d, 0)$ after the h_0-th iteration. Let $\boldsymbol{\beta}(d) = (\beta_{11}(d), \beta_{12}(d), \beta_{13}(d), \beta_{21}(d), \beta_{22}(d), \beta_{23}(d))^T$ be a 6×1 unknown parameter vector. To assess both Type I and II error rates at the voxel level, we selected a region-of-interest (ROI) with 64 points on the reference sphere. We set $\boldsymbol{\beta}(d) = \mathbf{0}_6$ across the whole sphere and then change $\beta_{12}(d)$ from 0 to 2 for all points d in ROI (Fig 2(b)). We test the hypotheses $H_0 : \beta_{12}(d) = 0$ and $H_1 : \beta_{12}(d) \neq 0$ across all voxels. We applied the MARM with $c_h = 1.25$, $S = 6$ and computed the p-values of $W_\mu(d, h)$ at each iteration. The 10,000 replications were used to estimate the rejection rate with the significance level $\alpha = 5\%$. For the test statistic $W_\mu(d, h)$, the Type I rejection rates outside the ROI were relatively accurate for all radius, while the statistical power for rejecting the null hypothesis in the ROI was significantly increased with the the radius h (Table 3.2).

We simulated additional imaging datasets to examine the accuracy of $W_\mu(d, h)$ at the level of an entire brain region. To further assess the power and account for testing multiple hypotheses, we added an additional ROI with 17 voxels, in which $\beta_{12}(d)$ was also set at 2 (Fig. 3.2(c)). To introduce spatial correlation in the simulated imaging data, we smoothed the simulated residual data $\{\epsilon_i(d) : d \in \mathcal{D}\}$ on the sphere using

Table 3.2: Simulation Study for $W_\mu(d,h)$: True average rejection rate for voxels inside the ROI and false average rejection rate for voxels outside the ROI were reported at 6 different bandwidths ($h_s = 1.25^s$ and $h_0 = 0$) and 3 different sample sizes ($n = 20, 40, 60$) at $\alpha = 5\%$. For each case, 10,000 simulated datasets were used.

	$n = 20$		$n = 40$		$n = 60$	
s	True	False	True	False	True	False
0	0.2963	0.1029	0.3457	0.0782	0.6543	0.0762
1	0.5432	0.1311	0.8395	0.1104	0.9259	0.1145
2	0.6049	0.1256	0.8889	0.0938	0.9753	0.0787
3	0.8272	0.1205	0.9506	0.0872	0.9630	0.0777
4	0.8272	0.1130	0.9506	0.0877	0.9753	0.0792
5	0.8642	0.1225	0.9136	0.0837	0.9753	0.0772

heat kernel smoothing with 16 iterations. We used the rejection threshold based on the false discovery rate (FDR) at a q value equal to 0.2. Based on this threshold, we calculated the average of the probabilities of rejecting each of the 81 (=17+64) points in the two ROIs as an estimate of the average power using 1,000 replications and then we calculated the average of probabilities of rejecting the points outside of the two ROIs as an estimate of the average type I error. For $W_\mu(d,h)$, our test procedure worked very well and significantly outperformed the voxel-wise approach. The average power dramatically increases in detecting the significant voxels in the two ROIs as the radius increases, while the average type I error rates outside of the two ROIs are relatively low (Table 3.3).

3.3.3 Simulation Studies Part III

Following the setup in Section 3.3.1., we simulated an additional dataset at all the $m = 4002$ points on the surface of a hippocampus for 50 subjects, except that six new ROIs were constructed as three sets of nested circles. In the first set of nested circles, $\beta_2(d)$ were set at 1 and 2 in the inner and outer circles with radius being 2 and 4, respectively. In the second set of nested circles, $\beta_2(d)$ were set at 0.6 and 0.8 in the inner and outer circles at radius being 3 and 5, respectively. In the third set of nested

Table 3.3: Simulation Study for $W_\mu(d,h)$: true average rejection rate for voxels inside the two ROIs and false average rejection rate for voxels outside the two ROIs were reported at 6 different bandwidths ($h_s = 1.25^s$ and $h_0 = 0$) and 3 different sample sizes ($n = 20, 40, 60$) at a FDR q value at 0.2. For each case, 1,000 simulated datasets were used.

	$n = 20$		$n = 40$		$n = 60$	
s	True	False	True	False	True	False
0	0.1612	0.0325	0.1801	0.0098	0.3522	0.0072
1	0.5798	0.0799	0.8194	0.0451	0.9428	0.0325
2	0.6220	0.0601	0.8577	0.0315	0.9652	0.0289
3	0.7436	0.0566	0.9207	0.0319	0.9819	0.0276
4	0.7274	0.0541	0.9109	0.0312	0.9759	0.0275
5	0.8104	0.0594	0.9457	0.0352	0.9848	0.0279

circles, $\beta_2(d)$ was set 0.4 and 0.6 in the inner and outer circles at radius being 3 and 4, respectively. The $\beta_2(d)$ outside of these six ROIs were always set at 0. We use MARM to calculate the parameter estimate of β at 6 different scales. It is obviously that the estimate is more precisely estimated at h_5 compared with at h_0 at different signal to noise ratios (Fig. 3.3 (a) and (c)). Similarly, the p-value map generated for testing $H_0 : \beta_2(d)$=0 at h_5 also performs much better than that at $h_0 = 0$ (Fig. 3.3(b) and (d)).

We simulated MRI images using a state-of-art imaging method (Xue *et al.*, 2006). We first selected T1-weighted MR brain images from a group of 12 subjects, whose ages were over 65, and then simulated the atrophy at both precentral gyrus and superior temporal gyrus of these MR images to obtain an atrophy group. All of these 24 images, including 12 original images and 12 images with simulated atrophy, are spatially normalized on a template space by HAMMER. We calculated the grey matter tissue density maps from the estimated deformation fields. We used simulated deformations and images with the known ground truth to demonstrate the superiority of the MARM over the voxel-wise approach. We applied the MARM with $c_h = 1.25$, $S = 6$ and computed the p-values of $W_\mu(d,h)$ 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

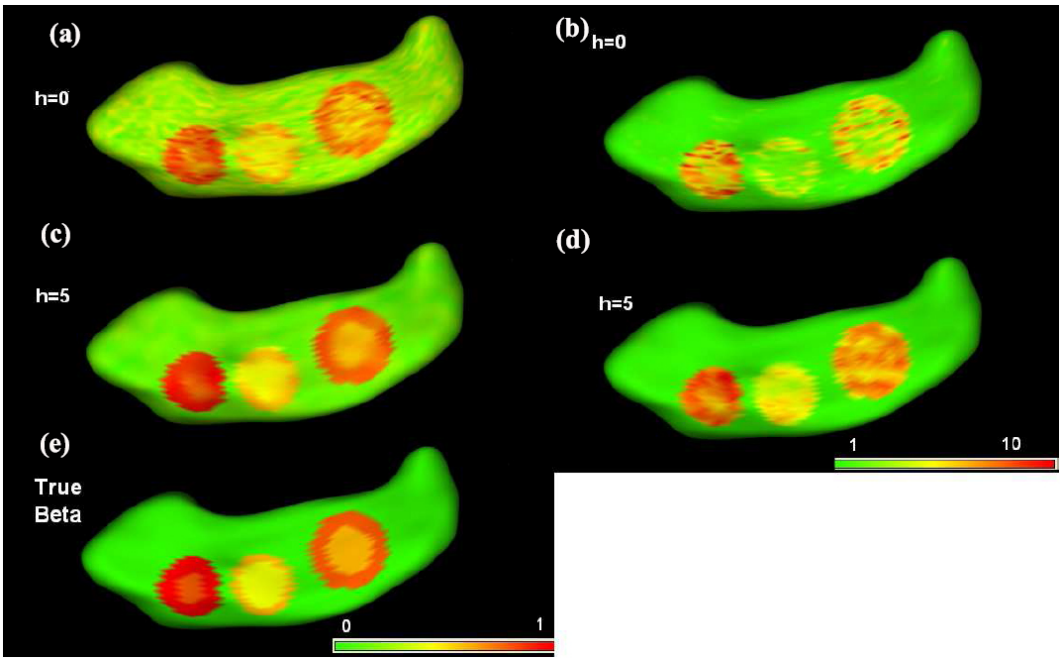

Figure 3.3: The maps of FDR corrected $-\log_{10}(p)$ values from two selected slices based on the voxel-wise approach (panels (a) and (c)) and MARM (panels (b) and (d)).

bandwidth h of the spherical neighborhood (Fig. 3.4 (b) and (d)). The MARM can correctly identify the simulated atrophy (Fig. 3.4 (b) and (d)), whereas the classical voxel-wise approach cannot (Fig. 3.4 (a) and (c)).

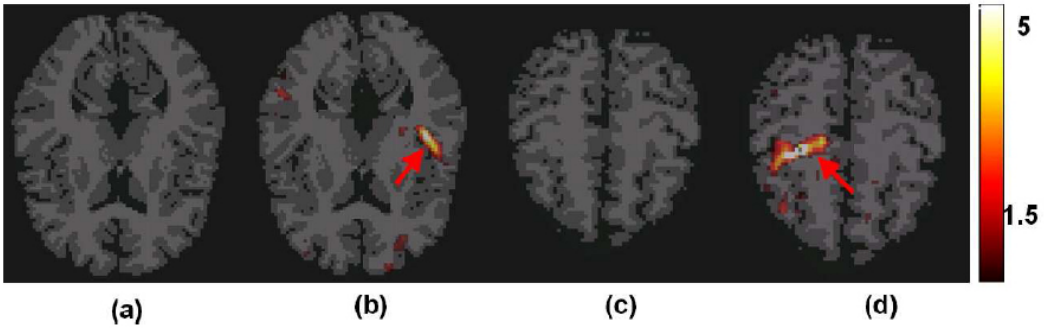

Figure 3.4: The maps of FDR corrected $-\log_{10}(p)$ values from two selected slices based on the voxel-wise approach (panels (a) and (c)) and MARM (panels (b) and (d)).

3.4 Real Data Analysis

Understanding white matter development in human brain *in vivo* is critical to the understanding of the functional formation of the central nervous system. An important feature of diffusion tensor imaging (DTI) is its capability in revealing white matter maturation process in human brain using a set of water diffusion related parameters, such as fractional anisotropy (FA) and radial (RD) diffusivity. For instance, FA represents the inhomogeneous extent of local barriers to water diffusion and has been widely used to investigate early brain development from identifying transient brain structures such as ganglionic eminance and cortical subpliate to estimating the correlation of white matter maturation with functional development measures such as IQ and working memory.

We considered 38 subjects from the neonatal project on early brain development led by Dr. Gilmore at the University of North Carolina at Chapel Hill. For each subject, diffusion-weighted images were acquired at 2 week, year 1 and year 2. Diffusion gradients with a $b-$value of 1000 s/mm^2 were applied in six non-collinear directions, (1,0,1), (-1,0,1), (0,1,1), (0,1,-1), (1,1,0), and (-1,1,0). A $b = 0$ reference scan was also obtained for diffusion tensor matrix calculations. Forty-six contiguous slices with a slice thickness of 2 mm covered a field of view (FOV) of 256×256 mm^2 with an isotropic voxel size of $2 \times 2 \times 2$ mm^3. Eighteen acquisitions were used to improve the signal-to-noise ratio (SNR) in the images. High resolution T1 weighted (T1W) images were acquired using a 3D MP-RAGE sequence. Then, a weighted least squares estimation method was used to construct the diffusion tensors (Basser, Mattiello, and LeBihan 1994 b; Zhu *et al.*, 2007b). All images were visually inspected before analysis to ensure no bulk motion. All DT images (38 subjects, 3 time points each) were registered, using TIMER, onto a randomly selected brain DT image of a 2-year-old. The aligned images were then voxel-wise averaged to create the mean DT image, from which the FA map can be computed (Yap *et al.*, 2009).

Fractional anisotropy (FA) calculated from DTIs is widely used as a measurement to

assess directional organization of the brain which is greatly influenced by the magnitude and orientation of white matter tracts. We use FA images to identify the spatial patterns of white matter maturation. We considered a linear model $y_i(d) = \beta_0(d) + t_i\beta_1(d) + t_i^2\beta_2(d) + \epsilon_i(d)$ for $i = 1, \cdots, n$, at each voxel of FA images. We applied the AET procedure with $c_h = 1.25$ and $S = 6$ to carry out statistical analysis and test $H_0 : \beta_1(d) = \beta_2(d) = 0$ for time effect across all voxels d. Compared with the results at h_0 (Fig. 3.5 (a)-(c)), MARM shows a clear advantage in detecting more significant and smooth activation areas as the bandwidth h increases (Fig. 3.5 (e)-(g)). In FA, internal capsule and corpus callosum including both splenium and genu have high FA values at birth. The linear coefficient in FA is positive in white matter region, while the quadratic coefficient in FA is mostly negative. Thus, a non-linear increasing pattern was observed for FA. We also selected two voxels with raw $-\log_{10}(p)$ values for $W_\mu(d, h_5)$ being 24.08 and 1.16 and plotted growth trajectories of FA values in these two voxels (Fig. 3.6 (a) and (b)).

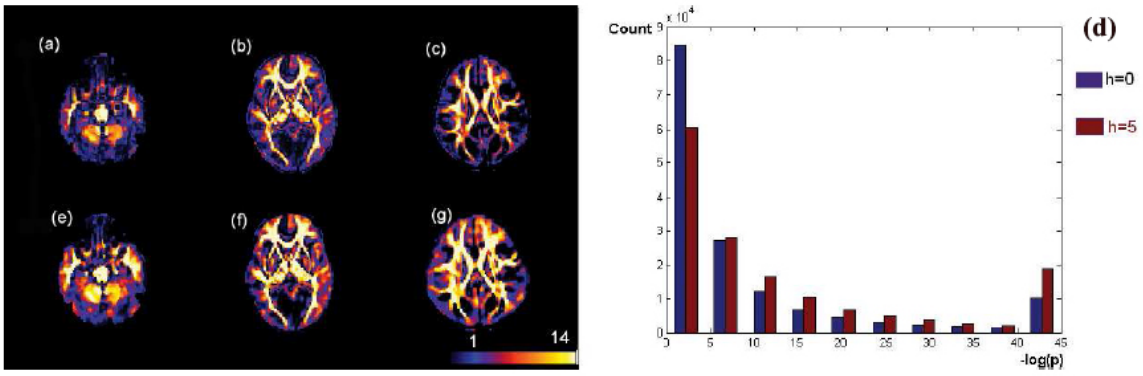

Figure 3.5: Results from the neonate project on brain development. Panels (a), (b) and (c): the raw $-\log_{10}(p)$ values of the Wald test statistics $W_\mu(d, h_0)$ from three selected slices; panels (e), (f) and (g): the raw $-\log_{10}(p)$ values of the Wald test statistics $\hat{W}_\mu(d, h_5)$ from the selected slices; (d) the comparison of the histograms for $W_\mu(d, h_0)$ and $W_\mu(d, h_5)$ across all voxels.

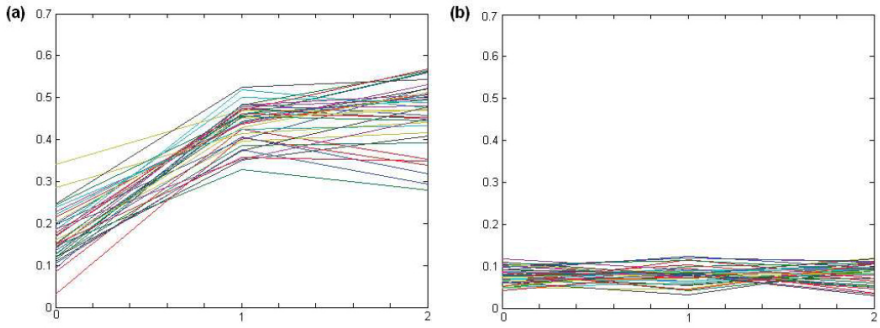

Figure 3.6: Growth trajectories of FA values in two selected voxels with the $-\log_{10}(p)$ values being: (a) $-\log_{10}(p) = 24.08$; (b) $-\log_{10}(p) = 1.16$.

3.5 Discussion

We have developed the MARM for spatial and adaptive analysis of imaging data. We have established consistency and asymptotic normality of the adaptive estimates and the asymptotic distributions of the adaptive test statistics. We have used simulation studies and real imaging data to demonstrate that the MARM significantly outperforms classical voxel-wise approach.

Many issues still merit further research. We will develop the multiscale method for generalized estimating equations, models with nonparametric component, and varying coefficient models and present them elsewhere.

3.6 Appendix

The following assumptions are needed to facilitate development of our methods, although they are not the weakest possible conditions.

(C1) $1 \geq \omega(d, d'; h) \geq 0$ and $\omega(d, d; h) = 1$ for all $d, d' \in \mathcal{D}$ and $h \geq 0$.

(C2) The data $\{\mathbf{Z}_i = (\mathbf{x}_i, \mathbf{Y}_{i,\mathcal{D}}) : i = 1, \cdots, n\}$ form an independent and identical sequence.

(C3) For any $d \in \mathcal{D}$, the maxima $\boldsymbol{\theta}_*(d)$ of $E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))]$ is an unique interior point of $\mathcal{B}$, where $\mathcal{B}$ is a compact set in R^p and the expectation is take with respect to

the true distribution of $Y(d)$ given $\mathbf{x}$.

(C4) For all voxels $d \in \mathcal{D}$, $\ell(\boldsymbol{\theta}(d)) = \log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))$ is twice continuously differentiable on Θ. For all $j, k, l = 1, \cdots, p$, $\ell(\boldsymbol{\theta}(d))$, $|\partial_j \ell(\boldsymbol{\theta}(d))|^2$, and $|\partial_j \partial_k \ell(\boldsymbol{\theta}(d))|^2$ are dominated by an integral function $G(Y(d), \mathbf{x})$ such that $E[\max_{d\in\mathcal{D}} |G(Y(d), \mathbf{x})|^r] < \infty$ for a $r > 1$, where $\partial_j = \partial/\partial\theta_j(d)$, in which $\theta_j(d)$ is the $j-$th component of $\boldsymbol{\theta}(d)$.

(C5) For a fixed $\delta > 0$,

$$\begin{aligned}
\infty &> \sup_{d\in\mathcal{D}} \max_{\boldsymbol{\theta}(d)\in B(\boldsymbol{\theta}_*(d),\delta)} (\lambda_{\max}\{E[-\partial^2_{\boldsymbol{\theta}(d)} \ell(\boldsymbol{\theta}(d))]\}) \\
&\geq \inf_{d\in\mathcal{D}} \min_{\boldsymbol{\theta}(d)\in B(\boldsymbol{\theta}_*(d),\delta)} (\lambda_{\min}\{E[-\partial^2_{\boldsymbol{\theta}(d)} \ell(\boldsymbol{\theta}(d))]\}) > 0, \\
\infty &> \sup_{d\in\mathcal{D}} \max_{\boldsymbol{\theta}(d)\in B(\boldsymbol{\theta}_*(d),\delta)} (\lambda_{\max}\{E[\partial_{\boldsymbol{\theta}(d)} \ell(\boldsymbol{\theta}(d))^{\otimes 2}]\}) \\
&\geq \inf_{d\in\mathcal{D}} \min_{\boldsymbol{\theta}(d)\in B(\boldsymbol{\theta}_*(d),\delta)} (\lambda_{\min}\{E[\partial_{\boldsymbol{\theta}(d)} \ell(\boldsymbol{\theta}(d))^{\otimes 2}]\}) > 0,
\end{aligned}$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix, respectively.

(C6) The kernel functions $K_{st}(u)$ and $K_{loc}(u)$ are continuous decreasing functions of $u \geq 0$ such that $K_{st}(0) = K_{loc}(0) = 1$, $\lim_{u\to\infty} K_{st}(u) = \lim_{u\to\infty} K_{loc}(u) = 0$, and $\lim_{u\to\infty} u^{1/2} K_{st}(u) = 0$.

(C7) $\lim_{n\to\infty} C_n/n = \lim_{n\to\infty} C_n^{-1} \log(N(\mathcal{D})) = \lim_{n\to\infty} C_n^{-1} = 0$.

Remarks A1: For each fixed $d \in \mathcal{D}$, Assumptions (C2)-(C5) are generalizations of the standard conditions for ensuring the first order asymptotic properties (e.g., consistency and asymptotic normality) of M-estimators (van der Vaart, 1998). Assumption (C2) is needed just for notational simplicity and can be easily modified to accommodate independent and non-identical distributed scenarios. Assumption (C3) is an identification condition, whereas Assumption (C4) is a uniform smoothness and integration condition. Particularly, Assumption (C4) ensures that $\ell(\boldsymbol{\theta}(d))$, $|\partial_j \ell(\boldsymbol{\theta}(d))|^2$, and $|\partial_k \partial_j \ell(\boldsymbol{\theta}(d))|^2$ are uniformly integrable for all $d \in \mathcal{D}$. Assumption (C5) is needed to ensure that the covariance matrix of $\hat{\boldsymbol{\theta}}(d, h)$ is positive definite for all $d \in \mathcal{D}$. Assumptions (C6) and (C7)

on $K_{st}(\cdot)$ and $K_{loc}(\cdot)$ is needed just for ensuring the desirable asymptotic properties of $\hat{\boldsymbol{\theta}}(d,h)$ and $W_\mu(d,h)$ based on the stochastic weights for the AET procedure.

Remarks A2: Assumption (C7) ensures that $\lim_{n\to\infty}\log(N(\mathcal{D}))/n = 0$. In neuroimaging data, although $N(\mathcal{D})$ is much larger than the sample size n, Assumption (C7) claims that we just need a relative large sample size compared with $\log(N(\mathcal{D}))$. For instance, in most neuroimaging data, $N(\mathcal{D}) \approx 100^3$ and $\log(10^3) = 14$. Therefore, a sample size such as 100 may be reasonable to use asymptotic normality to make statistical inference using MARM. Assumption (C7) is needed to invoke maximal inequalities (var der Vaart and Wellner, 1996). Moreover, Assumption (C7) also requires a large value of C_n relative to $\log N(\mathcal{D})$, but it may be weakened. In practice, we suggest to choose $C_n = n^\alpha$ for $\alpha \in (0,1)$.

Proof of Theorem 1. The proof of Theorem 1 consists of three steps. In Step 1, we will show that $\hat{\boldsymbol{\theta}}(h_0) = (\hat{\boldsymbol{\theta}}(d,h_0) : d \in \mathcal{D})$ converges $\boldsymbol{\theta}_* = (\hat{\boldsymbol{\theta}}_*(d) : d \in \mathcal{D})$ in probability. We need to introduce some notation. Let $\mathbf{T}$ be a bounded brain region in R^g containing all voxels $d \in \mathcal{D}$, where $g = 2$ for the 2D surface and $g = 3$ for the 3D volume. Let $\Theta = \prod_{d\in\mathcal{D}} \mathcal{B}$ be the parameter space for $\boldsymbol{\theta}$ and $\ell^\infty(\mathbf{T})^p$ is the product of p $\ell^\infty(\mathbf{T}) = \{z : \mathbf{T} \to R, \quad \sup_{\mathbf{t}\in\mathbf{T}} |z(\mathbf{t})| < \infty\}$. Let $\Psi_n : \Theta \to \ell^\infty(\mathbf{T})^p$ and $\Psi : \Theta \to \ell^\infty(\mathbf{T})^p$ be random maps and a deterministic map, respectively, such that

$$\Psi_n(\boldsymbol{\theta})(\mathbf{t}) = n^{-1}\sum_{i=1}^{n} \partial_{\boldsymbol{\theta}(d_\mathbf{t})} \log p(Y_i(d_\mathbf{t})|\mathbf{x}_i, \boldsymbol{\theta}(d_\mathbf{t})) \text{ and}$$

$$\Psi(\boldsymbol{\theta})(\mathbf{t}) = E[\partial_{\boldsymbol{\theta}(d_\mathbf{t})} \log p(Y(d_\mathbf{t})|\mathbf{x}, \boldsymbol{\theta}(d_\mathbf{t}))],$$

in which $d_\mathbf{t}$ denotes the voxel covering $\mathbf{t}$.

To prove the consistency of $\hat{\boldsymbol{\theta}}(h_0)$, we will show that

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\mathbf{t}\in\mathbf{T}} ||\Psi_n(\boldsymbol{\theta})(\mathbf{t}) - \Psi(\boldsymbol{\theta})(\mathbf{t})||_2 \to 0 \text{ and } \inf_{\boldsymbol{\theta}\in\Theta:||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\geq\epsilon} \sup_{\mathbf{t}\in\mathbf{T}} ||\Psi(\boldsymbol{\theta})(\mathbf{t})||_2 > \sup_{\mathbf{t}\in\mathbf{T}} ||\Psi(\boldsymbol{\theta}_*)(\mathbf{t})||_2. \tag{3.24}$$

It follows from Assumptions (C3) and (C4) that the second term in equation (4.22) is true. To prove the first term in equation (4.22), we note that

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{\mathbf{t}\in\mathbf{T}}||\Psi_n(\boldsymbol{\theta})(\mathbf{t})-\Psi(\boldsymbol{\theta})(\mathbf{t})||_2=\max_{d\in\mathcal{D}}A_n(d), \tag{3.25}$$

where $A_n(d)=\sup_{\boldsymbol{\theta}(d)\in\mathcal{B}}|n^{-1}\sum_{i=1}^{n}\partial_{\boldsymbol{\theta}(d)}\log p(Y_i(d)|\mathbf{x}_i,\boldsymbol{\theta}(d))-E[\partial_{\boldsymbol{\theta}(d)}\log p(Y(d)|\mathbf{x},\boldsymbol{\theta}(d))]|$. Then, we consider $\mathcal{F}=\{\partial_{\boldsymbol{\theta}(d)}\log p(Y(d)|\mathbf{x},\boldsymbol{\theta}(d)):d\in\mathcal{D},\boldsymbol{\theta}(d)\in\mathcal{B}\}$ with envelop $\max_{d\in\mathcal{D}}G(Y(d),\mathbf{x})$. Following the arguments in Theorem 2.4.3 of van der Vaart and Wellner (1996), we can show that $E[\max_{d\in\mathcal{D}}A_n(d)]$ is upper bounded by

$$\sqrt{[1+p\log(C_1(\epsilon)K)+\log(N(\mathcal{D}))]/n}C_2K$$
$$+2E[\max_{d\in\mathcal{D}}G(Y(d),\mathbf{x})\mathbf{1}\{\max_{d\in\mathcal{D}}G(Y(d),\mathbf{x})>K\}]+\epsilon\to 0,$$

where C_2 is a constant independent of ϵ, K can be chosen such that the second term of the above equation is arbitrarily small, and $C_1(\epsilon)$ is a constant depending on ϵ. Finally, following the arguments in Theorems 5.7 and 5.9 of van der Vaart (1998), we can prove the consistency of $\hat{\boldsymbol{\theta}}(h_0)$.

In Step 2, we will prove the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\theta}}(h_0)-\boldsymbol{\theta}_*)$. For each $d\in\mathcal{D}$, a Taylor expansion gives

$$\mathbf{0}=\Psi_n(\hat{\boldsymbol{\theta}}(h_0))(d)=\Psi_n(\boldsymbol{\theta}_*)(d)+\partial_{\boldsymbol{\theta}(d)}\Psi_n(\tilde{\boldsymbol{\theta}})(d)[\hat{\boldsymbol{\theta}}(d,h_0)-\boldsymbol{\theta}_*(d)], \tag{3.26}$$

where $\tilde{\boldsymbol{\theta}}\in\Theta$ and $\tilde{\boldsymbol{\theta}}(d)$ is on the line connecting $\boldsymbol{\theta}(d)$ and $\boldsymbol{\theta}_*(d)$. Similar to the proof of (4.23), we can show that

$$\sup_{\boldsymbol{\theta}\in\Theta:||\boldsymbol{\theta}-\boldsymbol{\theta}_*||_2\le\epsilon}\sup_{\mathbf{t}\in\mathbf{T}}||\partial_{\boldsymbol{\theta}(d_{\mathbf{t}})}\Psi_n(\boldsymbol{\theta})(\mathbf{t})-\partial_{\boldsymbol{\theta}(d_{\mathbf{t}})}\Psi(\boldsymbol{\theta})(\mathbf{t})||_2\to 0 \tag{3.27}$$

in probability, when $\log(N(\mathcal{D}))/n$ is sufficiently small. Therefore, we can show that

$$\sqrt{n}[\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)] = [-\partial_{\boldsymbol{\theta}_{(d)}} \Psi(\boldsymbol{\theta}_*)(d) + o_{p,\mathcal{D}}(1)]^{-1} \sqrt{n} \Psi_n(\boldsymbol{\theta}_*)(d), \tag{3.28}$$

for all $d \in \mathcal{D}$, where $o_{p,\mathcal{D}}(1)$ denotes the uniform convergence to zero for all $d \in \mathcal{D}$. It is easy to prove the asymptotic normality of $\sqrt{n}[\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)]$ for each $d \in \mathcal{D}$. Furthermore, by using Theorem 2.14.1 of van der Vaart and Wellner (1996), we can show that $\sup_{d \in \mathcal{D}} ||\Psi_n(\boldsymbol{\theta}_*)(d)||_2 = O_p(\sqrt{\log(N(\mathcal{D}))/n})$, which yields

$$\max_{d \in \mathcal{D}} ||\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)||_2 = O_p(\sqrt{\log N(\mathcal{D})/n}). \tag{3.29}$$

In Step 3, we will derive the rate of $D_{\boldsymbol{\theta}}(d, d'; h_0)$. Since $D_{\boldsymbol{\theta}}(d, d'; h_0)$ can be rewritten as

$$n[\hat{\Delta}(d, 0) - \hat{\Delta}(d', 0) + \triangle_*(d, d')]^T \Sigma_*(d, h)^{-1} [\hat{\Delta}(d, 0) - \hat{\Delta}(d', 0) + \triangle_*(d, d')][1 + o_p(1)],$$

it follows from (4.27) that if $\triangle_*(d, d') = \mathbf{0}$, then $\max_{d,d' \in \mathcal{D}} |D_{\boldsymbol{\theta}}(d, d'; h_0)| = O_p(\log(N(\mathcal{D})))$ and $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n) = K_{st}(O_p(\log(N(\mathcal{D})))/C_n) = 1 + o_p(1)$. However, if $\triangle_*(d, d') \neq \mathbf{0}$, then we have

$$D_{\boldsymbol{\theta}}(d, d'; h_0) = n||[\Sigma_*(d, h)]^{-1/2} [\triangle_*(d, d') + O_p(\sqrt{\log N(\mathcal{D})/n})]||_2^2,$$

which yields the proof of Theorem 1.

Proof of Theorem 2. We prove Theorem 2 (a) and (b) by induction. The proof primarily consists of three steps: (i) $s = 0$; (ii) $s = 1$; (iii) $s \geq 1$. In Step 1, we have already proved the case $s = 0$ in Theorem 1.

We prove Step 2 as follows. It follows from the definition of $\tilde{\omega}(d, d'; h_1)$ that

$$\sup_{\boldsymbol{\theta}(d)\in\mathcal{B}} |n^{-1}\ell_n(\boldsymbol{\theta}(d); h_1, \tilde{\boldsymbol{\omega}}) - M(\boldsymbol{\theta}(d); h_1, \tilde{\boldsymbol{\omega}})| \leq \sum_{d'\in B(d,h_1)} \tilde{\omega}(d, d'; h_1)\delta_n(d') \leq \max_{d'\in B(d,h_1)} \delta_n(d'),$$

where $\delta_n(d) = \sup_{\boldsymbol{\theta}(d)\in\mathcal{B}} |n^{-1}\sum_{i=1}^n \log p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d)) - E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))]|$. Then, following arguments in Theorems 2.7.11 and 2.4.3 of van der Vaart and Wellner (1996) and assumptions (C2)-(C4), we can show that

$$E[\max_{d\in\mathcal{D}} \delta_n(d)] \leq \sqrt{[1 + p\log(C_1(\epsilon)K) + \log(N(\mathcal{D}))]/n}C_2K +$$
$$2E[\max_{d\in\mathcal{D}} G(Y(d), \mathbf{x})\mathbf{1}\{\max_{d\in\mathcal{D}} G(Y(d), \mathbf{x}) > K\}] + \epsilon \to 0.$$

Since the above arguments are independent of $\tilde{\omega}(d, d'; h_1)$, we can conclude that

$$\max_{d\in\mathcal{D}} \sup_{\boldsymbol{\theta}(d)\in\mathcal{B}} |n^{-1}\ell_n(\boldsymbol{\theta}(d); h_1, \tilde{\boldsymbol{\omega}}) - M(\boldsymbol{\theta}(d); h_1, \tilde{\boldsymbol{\omega}})| \to 0 \tag{3.30}$$

in probability holds for any adaptive weights $\tilde{\omega}(d, d'; h)$.

Let $\mathcal{D}_*(d)^c = \{d' : \triangle_*(d, d') \neq \mathbf{0}\}$ and $\mathcal{D}_*(d) = \{d' : \triangle_*(d, d') = \mathbf{0}\}$. According to Theorem 1 (c), for all $d' \in B(d, h_1) \cap \mathcal{D}_*(d)^c$ and any $d \in \mathcal{D}$, we have

$$\begin{aligned} C_n^{-1} D_{\boldsymbol{\theta}}(d, d'; h_0) &= nC_n^{-1}\lambda_{\max}(\Sigma_*(d, h_0))^{-1} \times \\ &\quad \inf_{d'\in\mathcal{D}_*(d)^c} ||\triangle_*(d, d') + O_p(n^{-1/2})||_2^2 = \tilde{\delta}_n(d) \to \infty. \end{aligned} \tag{3.31}$$

It follows from (3.31) and (4.27) that

$$\begin{aligned} &\max_{d\in\mathcal{D}} \sup_{\boldsymbol{\theta}(d)} \left| M(\boldsymbol{\theta}(d); h_1, \tilde{\boldsymbol{\omega}}) - \sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \tilde{\omega}(d, d'; h) E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))] \right| \\ \leq\ &\max_{d\in\mathcal{D}} K_{st}(\tilde{\delta}_n(d)) E[\max_{d\in\mathcal{D}} G(Y(d), \mathbf{x})] \to 0. \end{aligned} \tag{3.32}$$

Since $\boldsymbol{\theta}_*(d) = \text{argmax}_{\boldsymbol{\theta}(d)} \sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \tilde{\omega}(d, d'; h) E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))]$, it follows from

Theorem 5.7 of var der Vaart (1998) and the arguments in the proof of Theorem 1 (a) that $\hat{\boldsymbol{\theta}}(h_1) = (\hat{\boldsymbol{\theta}}(d, h_1) : d \in \mathcal{D})$ converges to $\boldsymbol{\theta}_*$ in probability.

To prove the asymptotic normality of $\hat{\boldsymbol{\theta}}(d, h_1)$, we can use a Taylor expansion to show that

$$\mathbf{0} = \partial_{\boldsymbol{\theta}_{(d)}} \ell_n(\hat{\boldsymbol{\theta}}(d, h_1); h_1, \tilde{\boldsymbol{\omega}}) = \partial_{\boldsymbol{\theta}_{(d)}} \ell_n(\boldsymbol{\theta}_*(d); h_1, \tilde{\boldsymbol{\omega}}) + \partial^2_{\boldsymbol{\theta}_{(d)}} \ell_n(\tilde{\boldsymbol{\theta}}(d, h_1); h_1, \tilde{\boldsymbol{\omega}})[\hat{\boldsymbol{\theta}}(d, h_1) - \boldsymbol{\theta}_*(d)],$$

where $\tilde{\boldsymbol{\theta}}(d, h_1)$ is on the segment joining $\hat{\boldsymbol{\theta}}(d, h_1)$ and $\boldsymbol{\theta}_*(d)$. Similar to the Taylor's series expansion to show that and (4.29), we can show that

$$\begin{aligned}
&\max_{d \in \mathcal{D}} \sup_{||\boldsymbol{\theta}_*(d) - \boldsymbol{\theta}(d)||_2 \leq \epsilon} |n^{-1} \partial^2_{\boldsymbol{\theta}_{(d)}} \ell_n(\boldsymbol{\theta}(d); h_1, \tilde{\boldsymbol{\omega}}) - \\
&\quad \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) E[\partial^2_{\boldsymbol{\theta}_{(d)}} \log p(Y(d') | \mathbf{x}, \boldsymbol{\theta}(d))]| \to 0, \\
&\max_{d \in \mathcal{D}} n^{-1/2} |\partial_{\boldsymbol{\theta}_{(d)}} \ell_n(\boldsymbol{\theta}_*(d); h_1, \tilde{\boldsymbol{\omega}}) - \\
&\quad \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) \sum_{i=1}^{n} \partial_{\boldsymbol{\theta}_{(d)}} \log p(Y_i(d') | \mathbf{x}, \boldsymbol{\theta}_*(d))| \\
\leq \;& n^{1/2} K_{st} (O_p(n C_n^{-1})) E[\sup_{d \in \mathcal{D}} G(Y(d), \mathbf{x})] O(1) \to 0.
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
\sqrt{n}[\hat{\boldsymbol{\theta}}(d, h_1) - \boldsymbol{\theta}_*(d)] \;=\; & \{ - \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) E[\partial^2_{\boldsymbol{\theta}_{(d)}} \log p(Y(d') | \mathbf{x}, \boldsymbol{\theta}_*(d))] \\
& + o_{p, \mathcal{D}}(1) \}^{-1} \times n^{-1/2} \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) \\
& \sum_{i=1}^{n} \partial_{\boldsymbol{\theta}_{(d)}} \log p(Y_i(d') | \mathbf{x}, \boldsymbol{\theta}_*(d)). \qquad (3.33)
\end{aligned}$$

By using Theorem 2.14.1 of van der Vaart and Wellner (1996), we can show that

$$\max_{d \in \mathcal{D}} ||n^{-1/2} \sum_{i=1}^{n} \partial_{\boldsymbol{\theta}_{(d)}} \log p(Y_i(d') | \mathbf{x}, \boldsymbol{\theta}_*(d))||_2 = O_p(\sqrt{\log N(\mathcal{D})}),$$

which yields that $\max_{d\in\mathcal{D}} |||[\hat{\boldsymbol{\theta}}(d,h) - \boldsymbol{\theta}_*(d)]|||_2 = O_p(\sqrt{\log N(\mathcal{D})/n})$. Based on these results for $\hat{\boldsymbol{\theta}}(d,h_1)$, we can prove the same results as Theorem 1 (c) and (d) for $D_{\boldsymbol{\theta}}(d,d';h_1)$ and $K_{st}(D_{\boldsymbol{\theta}}(d,d';h_1)C_n^{-1})$.

In Step 3, by using the induction and the above arguments in Step 2, we can prove Theorem 2 (a) and (b) for any fixed $s > 1$.

Given the results in Theorem 2 (a) and (b), we can apply the standard arguments in the literature to prove Theorem 2 (c). We omit the details for simplicity.

Proof of Corollary 1. Because we can prove Corollary 1 (a) using the same arguments in proving Theorem 2 (a), we omit the details.

The proof of Corollary 1 (b) consists of two steps. In Step 1, following the same arguments in Theorem 2 (a), we can prove (4.30). In Step 2, we examine the asymptotic distribution of

$$A(d;h_1) = \sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \omega(d,d';h) n^{-1/2} \sum_{i=1}^{n} \partial_{\boldsymbol{\theta}_{(d)}} \log p(Y_i(d')|\mathbf{x},\boldsymbol{\theta}(d)_*).$$

For any $d' \in B(d,h_1)\cap\mathcal{D}_*(d)$, $D_{\boldsymbol{\theta}}(d,d';h_0)$ converges to a random variable, denoted by $Z(d,d';h_0)$, in distribution, and thus $\omega(d,d';h)$ converges to $K_{st}(Z(d,d';h_0))$ in distribution. In addition, for any $d' \in B(d,h_1)\cap\mathcal{D}_*(d)$, $n^{-1/2}\sum_{i=1}^{n} \partial_{\boldsymbol{\theta}_{(d)}} \log p(Y_i(d')|\mathbf{x},\boldsymbol{\theta}(d)_*)$ converges to a normal random vector, denoted by $Z(d')$, in distribution. Note that $Z(d')$ and $Z(d,d';h_0)$ are correlated with each other. Finally, using the continuous mapping theorem, we can claim that $A(d;h_1)$ converges to

$$\sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} K_{loc}(||d-d'||_2/h_1) K_{st}(Z(d,d';h_0))Z(d),$$

which is not a normal random variable when there is a $d' \in B(d,h_1)\cap\mathcal{D}_*(d)$. Thus, $W_\mu(d,h_1)$ is not asymptotically χ^2 distributed.

Proof of Theorem 3. We prove Theorem 3 (a) using induction. The proof primarily consists of two steps: (i) $\sqrt{n}[\hat{\boldsymbol{\beta}}(d,h_0) - \boldsymbol{\beta}_*(d)] = A_1(d;h_0)$ in probability; (ii) $\sqrt{n}[\hat{\boldsymbol{\beta}}(d,h_1) -$

$\boldsymbol{\beta}_*(d)] = A_1(d;h_1) + o_p(1)$ for each voxel d. Moreover, for notational simplicity, we assume that $\tau(d)$ are known through the proof.

In Step 1, since $\hat{\boldsymbol{\beta}}(d,h_0) = (\sum_{i=1}^n \mathbf{x}_i^{\otimes 2})^{-1} \sum_{i=1}^n \mathbf{x}_i Y_i(d) = \boldsymbol{\beta}_*(d) + A_1(d;h_0)/\sqrt{n} = \boldsymbol{\beta}_*(d) + (\sum_{i=1}^n \mathbf{x}_i^{\otimes 2})^{-1} \sum_{i=1}^n \mathbf{x}_i \epsilon_i(d)$ holds and $A_1(d;h_0)$ converges to $E[\mathbf{x}^{\otimes 2}]^{-1/2} Z(d)$ in distribution for any voxel d. Following the arguments in Theorem 2.4.3, we can show that $\max_{d\in\mathcal{D}} ||n^{-1}\sum_{i=1}^n \mathbf{x}_i\epsilon_i(d)||_2 = O_p(\sqrt{\log(N(\mathcal{D}))/n})$.

In Step 2, since $D_{\boldsymbol{\beta}}(d,d';h_0)$ can be rewritten as

$$n\tau(d)||E[\mathbf{x}^{\otimes 2}]^{-1/2}\{\triangle_*(d,d') + (\sum_{i=1}^n \mathbf{x}_i^{\otimes 2})^{-1} \sum_{i=1}^n \mathbf{x}_i[\epsilon_i(d') - \epsilon_i(d)]\}||_2^2,$$

where $\triangle_*(d,d') = \boldsymbol{\beta}_*(d) - \boldsymbol{\beta}_*(d')$, we can check that $D_{\boldsymbol{\beta}}(d,d';h_0)$ and $K_{st}(D_{\boldsymbol{\beta}}(d,d';h_0)/C_n)$ have the asymptotic expansions as described in Lemma 1. We can show that $\tilde{\omega}(d,d';h_1)$ are smaller than $K_{st}(O_p(nC_n^{-1}))$ for all $d' \in B(d,h_1) \cap \mathcal{D}_*(d)^c$ and $\hat{\omega}(d,d';h_1)$ converges to $C(d,d';h_1)$ for all $d' \in B(d,h_1) \cap \mathcal{D}_*(d)$. Therefore, we have

$$\begin{aligned}\sqrt{n}[\hat{\boldsymbol{\beta}}(d,h_1) - \boldsymbol{\beta}_*(d)] &= \sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \tilde{\omega}(d,d';h_1) E[\mathbf{x}^{\otimes 2}]^{-1/2} n^{-1/2} \sum_{i=1}^n \mathbf{x}_i\epsilon_i(d') + o_p(1) \\ &= A_1(d,h) + o_p(1).\end{aligned}$$

Applying the continuous mapping theorem yields the weak convergence of $A_1(d,h_1)$ and $\sqrt{n}[\hat{\boldsymbol{\beta}}(d,h_1) - \boldsymbol{\beta}_*(d)]$. We can use the same arguments in Corollary 1 (b) to prove Theorem 3 (b). Note that for the PS approach, $\hat{\omega}(d,d';h_1)$ converges in distribution to $C(d,d';h_1)K_{st}(\tau(d)||Z(d) - Z(d')||_2^2)$ for all $d' \in B(d,h_1) \cap \mathcal{D}_*(d)$.

Chapter 4

Multiscale Adaptive Generalized Estimating Equations for Longitudinal Neuroimaging Data

4.1 Introduction

Longitudinal imaging studies have been valuable for better understanding disease progression and normal brain development/aging. Compared to cross-sectional imaging studies, longitudinal imaging studies can increase the statistical power in detecting subtle spatiotemporal changes of brain structure and function.

The existing voxel analysis of neuroimaging data is sequentially executed in two stages. The first stage is a model fitting stage. It fits a general linear model or a simple linear mixed model to the data from all subjects at each voxel. The second stage is a multiple testing stage. It generates a statistical parametric map that contains a statistic (or a p-value) at each voxel (Worsley et al., 2004; Friston, 2007; Lau et al., 2008). The general linear model used in the neuroimaging literature usually involves two key assumptions: that the variance of the imaging data is homogeneous across subjects and that the data conform to a Gaussian distribution at each voxel. These two assumptions

are important for the valid calculation of parametric distributions in conventional tests (e.g., F test) that assess the statistical significance of parameter estimates in general linear model. It has been well known in the neuroimaging literature that the distribution of the univariate (or multivariate) neuroimaging measurements often deviates from the Gaussian distribution (Ashburner and Friston, 2000; Luo and Nichols, 2003; Zhu *et al.*, 2007a).

The existing voxel-wise methods have major limitations for analyzing neuroimaging data. (i) Spatial smoothing is commonly applied to all kinds of real imaging data including functional magnetic resonance images and diffusion tensor images prior to the formal model fitting stage. It has been well-known that the final results of voxel-based analysis can strongly depend on the amount of smoothing in the smoothed imaging data (Jones et al., 2005; Scouten et al., 2006; Weibull et al., 2008). The use of the common Gaussian kernel with arbitrary bandwidth for smoothing imaging data can blur the image data near the edges of the activated regions and subsequently, it will dramatically increase the numbers of false positives and negatives (Jones et al., 2005; Tabelow *et al.*, 2006). (ii) All voxel-wise approaches suffer from misalignment problem. That is, even after an image warping procedure, the location of a voxel in the image of one person is not in precisely the same location as the voxel identified in another person. Spatial smoothing real imaging data may potentially reduce the effect of imaging misalignment on final group analysis. (iii) The voxel-wise methods essentially treat all voxels as independent units in the model fitting stage (Tabelow *et al.*, 2006). In contrast, neuroimaging data are anticipated to contain spatially contiguous regions of activation with rather sharp edges.

The aims of this article are to develop a multiscale adaptive generalized estimating equation (MAGEE) for the spatial and adaptive analysis of longitudinal neuroimaging data and to demonstrate its superiority over the voxel-wise approach using simulated

and real imaging data. There are four features for MAGEE: being spatial, being semi-parametric, being hierarchical and being adaptive. MAGEE explicitly utilizes the spatial information to carry out statistical inference by constructing nested spheres with increasing radius at all voxels. Then, instead using any parametric distribution, MAGEE uses weighted generalized estimating equations to fit all observations in the voxels within the sphere of the current voxel. MAGEE constructs hierarchically nested spheres in adaptively computing parameter estimates and testing statistics. Thus, MAGEE can adaptively utilize available information in the neighboring voxels to increase the precision of parameter estimates and the power of test statistics in detecting subtle changes of brain structure and function.

MAGEE represents a novel generalization of the standard spatial smoothing techniques and the voxel-wise statistical methods for the analysis of longitudinal imaging data. The standard voxel-wise methods are sequentially executed in two independent steps: a smoothing step using an 'arbitrary' bandwidth and a statistical analysis step. In contrast, MAGEE is a simultaneous smoothing and estimation method, allowing adaptively smoothing images while accounting for the spatial pattern of activation regions. Most spatial smoothing techniques directly smoothing raw images from each subject at each time point independently. In contrast, MAGEE simultaneously smooth all raw images from all subjects across all time points using the learned information during the statistical estimation step. More importantly, instead of smoothing raw images, MAGEE can smooth images of all parameters of interest, while fixing images of other nuisance parameters. For instance, the scientific interests of many neuroimaging studies typically focus on the comparison of full tensors across groups, while controlling for age, gender, and other covariates of interest. MAGEE allows solely smoothing the image of diagnostic effect without distorting the images associated with other covariates, such as age and gender.

Compared with the Gaussian distributional assumption in the general linear model,

MAGEE is a semiparametric method and explicitly account for the temporal correlation existed between the repeated measurements from the same subject. Thus, it is very desirable for the analysis of longitudinal neuroimaging data. MAGEE also includes specific methods for approximating the standard errors of the smoothed parametric estimates. We also theoretically examine the adaptive weights in the MAGEE and their roles in ensuring the proper statistical properties of parameter estimators. Finally, we formalize some technical conditions and formally establish the asymptotic properties including consistency and asymptotic distributions of the parameter estimates and test statistics for MAGEE.

In Section 4.2 of this paper, we will present MAGEE just described and establish the associated theoretical properties. Particularly, we will establish consistency and asymptotic normality of the adaptive estimator and the asymptotic distribution of the adaptive score test statistic for MAGEE. In Section 4.3, we will conduct two sets of simulation studies with the known ground truth to examine the finite sample performance of MAGEE. Section 4.4 illustrates an application of MAGEE in a longitudinal DTI dataset acquired from 38 healthy full term unsedated babies at approximately two weeks, one year, and two years after birth. We present concluding remarks in Section 4.5.

4.2 Multiscale Adaptive Generalized Estimating Equations

4.2.1 Model Formulation

We observe imaging, behavioral and clinical data from n subjects in a longitudinal study. Let x_{ij} be a $q_x \times 1$ covariate vector of interest, which may include age, gender, height, and many others, for the i-th subject at the j-th time point t_{ij} for $i = 1, \cdots, n$ and $j = 1, \cdots, m_i$. Here m_i denotes the number of time points for the $i-$th subject and thus there are a total $\sum_{i=1}^{n} m_i = N$ images in the longitudinal study. For instance, in our

longitudinal study, we acquired the anatomical magnetic resonance images and diffusion weighted images of one cohort of neonates at birth, around one and two years old. Based on registered image data on the template, we can obtain neuroimaging measures from the ith subject, denoted by $Y_i = \{y_{ij}(d) : d \in \mathcal{D}, j = 1, \cdots, m_i\}$, where $y_{ij}(d)$ is a $p \times 1$ vector and d represents a voxel on the template $\mathcal{D}$. The dimension of $y_{ij}(d)$ can be either univariate or multivariate. For example, the spherical harmonic shape description (SPHARM) of subcortical surfaces is a set of three dimensional imaging measures across the subcortical surfaces (Styner and Gerig, 2003).

4.2.2 Voxel-wise Generalized Estimating Equations

We apply the generalized estimating equation (GEE) approach for jointly modeling multivariate (or univariate) imaging measures with behavioral and clinical variables at each voxel in longitudinal study settings (Liang and Zeger, 1986; Diggles et al., 2002). The GEE method as a semiparametric method is free of distributional assumption. Moreover, even under the misspecified correlation structure, the GEE estimators of regression parameters are consistent and the covariate matrix of the regression parameters can be consistently estimated using a sandwich estimator.

For simplicity, we temporarily drop voxel d from our notation. At a voxel d on the brain subregion, we consider the moments model

$$E(y_{ij}) = \mu_{ij} = \mu(x_{ij}, \beta) \qquad \text{for } \; i = 1, \cdots, n; \quad j = 1, \cdots, m_i, \tag{4.1}$$

where β is a $q \times 1$ vector and $\mu(\cdot, \cdot)$ is a $p \times 1$ vector of known monotonic functions, called link functions. Furthermore, we assume that the covariance matrix of $Y_i = (y_{i1}^T, \cdots, y_{im_i}^T)^T$ is given by

$$V_i(\theta) = A_i^{1/2}(\beta, \gamma)[R_{m_i}(\alpha_1) \otimes R_{p,i}(\alpha_2)]A_i^{1/2}(\beta, \gamma) \quad \text{for } \; i = 1, \cdots, n, \tag{4.2}$$

where $\theta = (\alpha, \beta, \gamma)$, $R_{m_i}(\alpha_1)$ and $R_P(\alpha_2)$, respectively, represent the correlation among all m_i repeated measurements over time and the correlation among all p imaging measures and $\alpha = (\alpha_1, \alpha_2)$ characterizes the unknown parameters in the correlation matrices. In addition, $A_i^{1/2}(\beta, \gamma)$ is a $pm_i \times pm_i$ diagonal matrix and contains the standard deviations of all pm_i measurements, where γ is the additional parameter vector for characterizing the variances of imaging measures.

If α and γ are known, then the GEE for β is given by

$$\sum_{i=1}^{n} D_i(\beta)^T A_i^{-1/2}(\beta, \gamma)[R_{m_i}(\alpha_1)^{-1} \otimes R_{p,i}(\alpha_2)^{-1}] A_i^{-1/2}(\beta, \gamma)[Y_i - \mu_i(\beta)] = 0, \qquad (4.3)$$

where $D_i(\beta) = \partial \mu_i(\beta)/\partial \beta$ is a $m_i p \times q$ matrix and $\mu_i(\beta) = (\mu_{i1}^T, \cdots, \mu_{im_i}^T)^T$. If $V_i(\theta)$ is correctly specified, then the GEE for β is Godambe efficient for estimating β (Godambe, 1960). However, α and γ are unknown and must be estimated. It is common to set up additional estimation equations for estimating α and γ. For instance, similar to Ye and Pan (2006), we may construct a set of estimating equations for estimating γ as follows:

$$\sum_{i=1}^{n} \frac{\partial \text{diag}(A_i(\beta, \gamma))}{\partial \gamma^T} W_i[\text{diag}((Y_i - \mu_i(\beta))^{\otimes 2}) - \text{diag}(A_i(\beta, \gamma))] = 0, \qquad (4.4)$$

where $a^{\otimes 2} = aa^T$ for any vector a and $\text{diag}(A)$ denotes the vector of all diagonal elements of matrix A. Moreover, W_i is a prespecified weighted matrix. For instance, we may set W_i to be an identity matrix to avoid making additional assumptions on the fourth moments of imaging measures. Numerically, we can resort to the Newton-Raphson algorithm to solve $\hat{\theta}$.

Generally, it is difficult to correctly specify $R_{m_i}(\alpha_1)$ and $R_{p,i}(\alpha_2)$, called working correlation matrices, in the GEE setting. Common used working correlation structures include $m-$dependent, exchangeable, autoregressive AR(1), and unstructured (Diggle et al., 2002). Under a selected working correlation structure, estimation procedures for estimating α_1 and α_2 can be constructed using the Pearson residuals. Even under the

misspecified correlation structures, the estimator of β for (4.1), denoted by $\hat{\beta}$, can be consistent and asymptotically normal (Liang and Zeger, 1986). Assume that $\hat{\theta}$ includes $\hat{\beta}$ and an estimator of (α, γ), denoted by $(\hat{\alpha}, \hat{\gamma})$. The covariance matrix of $\hat{\beta}$ can be approximated by

$$[\sum_{i=1}^{n} \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i]^{-1} [\sum_{i=1}^{n} \hat{D}_i^T \hat{V}_i^{-1} (Y_i - \hat{\mu}_i)^{\otimes 2} \hat{V}_i^{-1} \hat{D}_i] [\sum_{i=1}^{n} \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i]^{-1}, \tag{4.5}$$

where $\hat{D}_i = D_i(\hat{\beta})$, $\hat{V}_i = V_i(\hat{\theta})$, and $\hat{\mu}_i = \mu_i(\hat{\beta})$.

4.2.3 Weighted Generalized Estimating Equations

Neuroimaging data often contain spatially contiguous regions of activation with rather sharp edges, but the voxel-wise approach does not account for such spatial structure in neuroimaging data, which can lead to great loss of power in detecting statistical significance in the analysis of neuroimaging data. We propose a weighted GEE as a possible solution for accounting for the spatial structure in neuroimaging data as follows.

In longitudinal studies, $\beta(d)$ is the parameter of interest, while $(\alpha(d), \gamma(d))$ may be regarded as nuisance parameters. In neuroimaging data, voxels, which are not on the boundary of regions of activation, often have a neighborhood in which $\beta(d)$ is nearly constant. Thus, we may combine the GEEs for $\beta(d)$ in a neighboring sphere of d to make inference on $\beta(d)$. Specifically, let $B(d, h)$ be a sphere of d with the radius h, we introduce a weighted GEE, denoted by $G_n(\beta(d); \omega, h)$, which is defined as follows:

$$\sum_{i=1}^{n} \sum_{d' \in B(d,h)} \omega(d, d'; h) D_i(\beta(d))^T V_i(\alpha(d'), \beta(d), \gamma(d'))^{-1} [Y_i(d') - \mu_i(\beta(d))] = 0, \tag{4.6}$$

where $\omega(d, d'; h)$ characterizes the similarity between the data in voxels d and d'. Moreover, as detailed below, $\omega(d, d'; h)$ can prevent incorporating voxels whose data do not contain information on $\beta(d)$, and thus preserve the edges of the regions of activation. We require that $\omega(d, d'; r_0)$ be independent of i just for simplicity. Note that the weighted

GEE utilizes all data in $\{Y_i(d') : d' \in B(d,h)\}$.

We present the estimation method and test statistic based on the weighted GEE at each $d \in \mathcal{D}$ for a fixed radius h. Specifically, given the current weights $\{\omega(d,d';h) : d,d' \in \mathcal{D}\}$, we consider the weighted GEE estimator of $\beta(d)$, denoted by $\hat{\beta}(d,h)$, which is the solution of $G_n(\hat{\beta}(d,h);\omega,h) = 0$. It should be noted that $G_n(\hat{\beta}(d,h);\omega,h)$ contains nuisance parameters $\{(\alpha(d'),\gamma(d')) : d' \in B(d,h)\}$, but these nuisance parameters have negligible effects on the asymptotic distribution of $\hat{\beta}(d,h)$. Specifically, let $\hat{D}_i(d,h) = D_i(\hat{\beta}(d,h))$, $\hat{V}_{i,\omega}(d,h)^{-1} = \sum_{d' \in B(d,h)} \omega(d,d';h) V_i(\hat{\alpha}(d'),\hat{\beta}(d,h),\hat{\gamma}(d'))^{-1}$, and $e_i(d',\hat{\beta}(d,h)) = Y_i(d') - \mu_i(\hat{\beta}(d,h))$. The covariance matrix of $\hat{\beta}(d,h)$ can be approximated by

$$\mathrm{Cov}(\hat{\beta}(d,h)) \approx C_0(d,h)^{-1} C(d,h) C_0(d,h)^{-1}, \tag{4.7}$$

where $C_0(d,h) = \sum_{i=1}^n \hat{D}_i(d,h)^T \hat{V}_{i,\omega}(d,h)^{-1} \hat{D}_i(d,h)$ and

$$C(d,h) = \sum_{i=1}^n \hat{D}_i(d,h)^T [\sum_{d' \in B(d,h)} \omega(d,d';h) V_i(\alpha(d'),\beta(d),\gamma(d'))^{-1} e_i(d',\hat{\beta}(d,h))]^{\otimes 2} \hat{D}_i(d,h).$$

We develop the score test statistic for testing hypothesis of interest. In neuroimaging studies, the primary hypotheses of interest include a comparison of brain structure (or function) across diagnostic groups or the detection of a spatialtemporal change in brain structure (or function) (Styner *et al.*, 2005; Thompson and Toga, 2002; Zhu *et al.*, 2007a). Without loss of generality, we assume that these hypotheses can be formulated as follows:

$$H_0(d) : R\beta(d) = b_0 \quad \text{vs.} \quad H_1(d) : R\beta(d) \neq b_0, \tag{4.8}$$

where R is a $r \times k$ matrix of full row rank and b_0 is a $r \times 1$ specified vector.

We consider a score test statistic $S_W(d,h)$ defined by

$$S_W(d,h) = U_W(d,h)^T \hat{I}_W(d,h)^{-1} U_W(d,h) \tag{4.9}$$

where $U_W(d,h) = \sum_{i=1}^{n} \hat{U}_{i,w}(\tilde{\beta}(d,h))$ and $\hat{I}_W(d,h) = \sum_{i=1}^{n} \hat{U}_{i,w}(\tilde{\beta}(d,h))^{\otimes 2}$, in which $\tilde{\beta}(d,h)$ denotes the estimate of β under $H_0(d)$ and the explicit expressions of $\hat{U}_{i,w}(\tilde{\beta}(d,h))$ and $U_W(d,h)$ are given in Appendix. As shown in Theorem 1, under $H_0(d)$ and some mild conditions, $S_W(d,h)$ is asymptotically distributed as $\chi^2(r)$, a chi-square distribution with r degrees of freedom. To test whether $H_0(d)$ holds in all voxels of the region under study, we may consider resampling method, false discovery rate (FDR) method and the random field theory (Benjamini and Hochberg, 1995; Worsley *et al.*, 2004; Zhu et al., 2008).

4.2.4 Adaptive Estimation and Testing Procedure

We develop an adaptive estimation and testing procedure (AET) to determine $\{\omega(d,d';r_0) : d,d' \in \mathcal{D}\}$ and then we adaptively estimate $\beta(d)$ based on the weighted GEEs. At each $d \in \mathcal{D}$, the AET procedure evolves along a sequence of nested spheres with increasing radiuses $h_0 = 0 < h_1 < \cdots < h_S = r_0$ (panel (a) in Fig. 4.1). At the scale $h_0 = 0$, we use the voxel-wise GEE to estimate $\hat{\theta}(d) = (\hat{\alpha}(d), \hat{\beta}(d), \hat{\gamma}(d))$. This is equivalent to estimating $\theta(d)$ in the weighted GEE with fixing $\omega(d,d';h_0) = 1(d = d')$, in which $1(A)$ is an indicator of an event A. We then fix $(\alpha(d), \gamma(d))$ at $(\hat{\alpha}(d), \hat{\gamma}(d))$ and combine all information contained in $\{\hat{\beta}(d) : d \in \mathcal{D}\}$ to calculate weights $\omega(d,d';h_1)$ at scale h_1 for all $d \in \mathcal{D}$. Subsequently, we utilize all data in $\{B(d,h_1) : d \in \mathcal{D}\}$, all weights $\{\omega(d,d';h_1) : d,d' \in \mathcal{D}\}$, and the weighted GEEs to estimate $\hat{\beta}(d;h_1)$ across $\mathcal{D}$. In this way, we can sequentially determine $\omega(d,d';h_s)$ and adaptively update $\hat{\beta}(d;h_s)$ from $h_0 = 0$ to $h_S = r_0$. The key feature of the AET method is to gradually smooth the images of parameter estimates, which can also decrease the variability of the calculated weights. At the end of AET, we calculate the score test statistics across all voxels. A

path diagram of the AET procedure is given as follows:

$$\begin{array}{ccccccc} \omega(d,d';h_0) & & \omega(d,d';h_1) & & \cdots & & \omega(d,d';h_S=r_0) \\ \Downarrow & \nearrow & \Downarrow & \nearrow & \cdots & \nearrow & \Downarrow \\ \hat{\theta}(d) & & \hat{\beta}(d;h_1) & & \cdots & & (\hat{\beta}(d;h_S), S_W(d,h_S)). \end{array} \tag{4.10}$$

The AET procedure has four key steps: initialization, weights adaptation, estimation, and stopping. In the initialization step (i), we prefix a geometric series $\{h_s = c_h^s : s = 1, \cdots, S\}$ of radiuses with $h_0 = 0$, where $c_h \in (1, 2)$, say $c_h = 1.5$. At each voxel d, we calculate the GEE estimate $\hat{\theta}(d) = (\hat{\alpha}(d), \hat{\beta}(d), \hat{\gamma}(d))$. From now on, we fix $(\alpha(d), \gamma(d))$ at $(\hat{\alpha}(d), \hat{\gamma}(d))$. We then set $s = 1$ and $h_1 = c_h$.

In the weight adaptation step (ii), we calculate the adaptive weights as follows:

$$\omega(d, d'; h_s) = K_{loc}(||d - d'||_2/h_s) K_{st}(d, d'; n, h_{s-1}), \tag{4.11}$$

where $K_{loc}(u)$ is a regular kernel function for smoothing curves or surfaces and $||\cdot||_2$ denotes the Euclidean norm of a vector (or a matrix). The weight $K_{loc}(||d - d'||_2/h_s)$ gives less weight to the voxel $d' \in B(d, h_s)$, whose location is far from the voxel d. Some common choices of $K_{loc}(u)$ include the Gaussian kernel and the Epanechnikov kernel (Fan and Gijbels, 1996; Tabelow *et al.*, 2006; Polzehl and Spokoiny, 2000, 2006). Without loss of generality, we use $K_{loc}(u) = (1 - u)_+$. Moreover, $K_{st}(d, d'; n, h)$ is a function of voxels d and d', the sample size n and the radius h. The adaptive weight $K_{st}(d, d'; n, h_{s-1})$ downweights the role of a voxel $d' \in B(d, h_s)$ in the weighted GEE $G_n(\beta; \omega, h) = 0$ in (4.6), if the data in voxel d differs substantially from those in voxel d'. Specific choices of $K_{st}(d, d'; n, h)$ will be detailed later.

In the estimation step (iii), for the radius h_s, we calculate the weighted GEE estimator $\hat{\beta}(d, h_s)$, which is defined in (4.11), at each voxel $d \in \mathcal{D}$.

In the stopping step (iv), when $s = S$, we compute $\hat{\beta}(d; h_S)$ and the p-values for

$S_W(d, h_S)$, apply either FDR or RFT to detect significant voxels and then stop. Otherwise, we 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 to the whole volume.

Remark 1. The additional computational time for MAGEE is very light compared to the voxel-wise approach. At the s-th iteration, we always set the final estimator $\hat{\beta}(d, h_{s-1})$ from the $(s-1)$-th iteration as the initial value $\hat{\beta}(d, h_s)^{(0)}$ in the Newton-Raphson algorithm. Since the AET procedure always downweights the data in voxel $d' \in B(d, h)$ when the data in voxel d' differs substantially from those in voxel d, $\hat{\beta}(d, h_{s-1})$ and $\hat{\beta}(d, h_s)$ should be close to each other. By starting from $\hat{\beta}(d, h_{s-1})$, the Newton-Raphson algorithm for the s-th iteration converges very fast.

Remark 2. The $K_{st}(d, d'; n, h)$ is a kernel function for downweighting voxel d', whose feature is dissimilar to that of voxel d during the process of making inference on $\beta(d)$. A particular choice of $K_{st}(d, d'; n, h)$ is a weighted distance between $\hat{\beta}(d, h_{s-1})$ and $\hat{\beta}(d', h_{s-1})$ as follows:

$$K_{st}(d, d'; n, h) = e^{-D_\beta(d,d';h_{s-1})/C_n} 1(D_\beta(d, d'; h_{s-1})/C_n \leq C_0) \tag{4.12}$$

where C_n is a scalar associated with n, C_0 is a prefixed number, and $D_\beta(d, d'; h_{s-1})$ is defined by

$$[\hat{\beta}(d, h_{s-1}) - \hat{\beta}(d', h_{s-1})]^T \mathrm{Cov}(\hat{\beta}(d; h_{s-1}))^{-1} [\hat{\beta}(d, h_{s-1}) - \hat{\beta}(d', h_{s-1})]. \tag{4.13}$$

We may select C_n as the logarithm of the number of voxels in $B(d, h)$ and the quantile of the χ^2 distribution (Polzehl and Spokoiny 2000, 2006). Note that $K_{st}(d, d'; n, h)$ in (4.13) changes across iterations.

We may consider another choice of $K_{st}(d, d'; n, h)$ based on a multivariate signed-rank

test statistic as follows (Haataja et al., 2009). We consider the data from voxels d and d' and define a $p \times m_i$ matrix given by

$$Z_i(d, d') = (\Delta y_{i1}(d, d'), \cdots, \Delta y_{im_i}(d, d')), \quad i = 1, \cdots, n, \tag{4.14}$$

where $\Delta y_{ij}(d, d') = y_{ij}(d) - y_{ij}(d')$. Let the spatial sign of $\Delta y_{ij}(d, d')$, denoted by $S(\Delta y_{ij}(d, d'))$, equal $||y_j(d) - y_{ij}(d')||_2^{-1}[y_{ij}(d) - y_{ij}(d')]$ if $y_{ij}(d) - y_{ij}(d') \neq 0$ and 0 if $y_{ij}(d) - y_{ij}(d') = 0$. The multivariate signed-rank test statistic is given by

$$U(d, d') = \sum_{i=1}^{n} \sum_{j=1}^{m_i} Q(\Delta y_{ij}(d, d')), \tag{4.15}$$

where $Q(\Delta y_{ij}(d, d'))$ is the spatial signed-rank centered around 0 as follows:

$$0.5N^{-1} \sum_{r=1}^{n} \sum_{s=1}^{m_r} [S(\Delta y_{ij}(d, d') - \Delta y_{rs}(d, d')) - S(\Delta y_{ij}(d, d') + \Delta y_{rs}(d, d'))]. \tag{4.16}$$

We introduce a weighted distance of $U(d, d')$, denoted by $D_U(d, d')$, as follows:

$$D_U(d, d') = U(d, d')^T \{\sum_{i=1}^{n} [\sum_{j=1}^{m_i} Q(\Delta y_{ij}(d, d'))]^{\otimes 2}\}^{-1} U(d, d'). \tag{4.17}$$

If $\mu_i(\beta(d)) = \mu_i(\beta(d'))$, then it can be shown under some mild conditions that $D_U(d, d')$ converges to $\chi^2(p)$ in distribution as $n \to \infty$ (Haataja et al., 2009). For the unbalanced design, we may consider a weighted version of $D_U(d, d')$. Finally, we can use $D_U(d, d')$ to define $K_{st}(d, d'; n, h)$ as follows:

$$K_{st}(d, d'; n, h) = e^{-D_U(d,d')/C_n} 1(D_U(d, d')/C_n \leq C_0). \tag{4.18}$$

It should be noted that $K_{st}(d, d'; n, h)$ in (4.18) solely depends the data in voxels d and d' and does not change during iterations.

Remark 3. We have developed the AET procedure for solely smoothing all components of $\beta(d)$ in the 3D volume (or 2D surface). Because $\beta(d)$ is statistically 'orthogonal' to $(\alpha(d), \gamma(d))$, we can develop the above AET procedure without updating $(\alpha(d), \gamma(d))$ at each iteration. However, we can easily modify the AET procedure to simultaneously smooth all components of $\theta(d)$.

4.2.5 Theoretical Properties

Throughout the paper, we only consider the asymptotic properties of $\hat{\beta}(d, h_s)$ and $S_W(d; h_s)$ for a finite number of iterations and bounded r_0 for MAGEE. We assume that the number of voxels in the brain volume does not increase with the sample size, since the resolution of a given imaging dataset is always fixed.

We establish consistency and asymptotically normality of $\hat{\beta}(d, h)$ and $S_W(d; h)$ for each h obtained from the AET procedure in Section 4.2.2. We first discuss the case with fixed weights $\omega(d, d'; h)$ for a fixed scale h. According to (4.6), the WGEE estimator $\hat{\beta}(d, h)$ solves the equation $0 = n^{-1}G_n(\beta(d); \omega, h)$, which converges to

$$G(\beta(d); \omega, h) = \sum_{i=1}^{n} \sum_{d' \in B(d,h)} \omega(d, d'; h) E\{D_i(\beta(d))^T V_i(\beta(d), d')^{-1} e_i(d', \beta(d))\}, \quad (4.19)$$

in probability under some mild conditions as $n \to \infty$ (var der Vaart, 1998), where $V_i(\beta(d), d') = V_i(\alpha(d'), \beta(d), \gamma(d'))$ and the expectation is taken with respect to $\{(Y(d', h), x) : d' \in B(d, h)\}$. Under some identifiability conditions, $\hat{\beta}(d; h)$ converges to $\beta_*(d; h)$, which solve the equation $G(\beta(d); \omega, h) = 0$ (van der Vaart, 1998). When $h = 0$, $\beta_*(d; 0) = \beta_*(d)$ is the 'pseudo' true value in voxel d. When $h > 0$, $\beta_*(d; h)$ can only be regarded as a weighted combination of all $\beta_*(d')$ for $d' \in B(d, h)$. In a homogeneous region, that is $\beta_*(d') = \beta_*(d)$, $\beta_*(d; h) = \beta_*(d)$ even for $h > 0$. However, in a nonhomogeneous region, an arbitrary set of weights $\omega(d, d'; h)$ can lead to undesirable consequences, such as smoothing out the boundary of activated regions and reducing statistical power

in detecting activated regions.

We obtain the following theorems, whose detailed assumptions and proofs can be found in the Appendix. We can establish important theoretical results to characterize the nice behavior of $\hat{\beta}(d,h)$ and $S_W(d,h)$ from the weighted GEE as follows.

Theorem 1. Suppose that Assumptions (C1)-(C7) in the Appendix are true. We have the following results for MAGEE:

(a) $\hat{\beta}(d,h)$ *converges to* $\beta_*(d)$ *in probability;*

(b) $\{Cov(\hat{\beta}(d,h))\}^{-1/2}[\hat{\beta}(d,h) - \beta_*(d)] \rightarrow^L N(0, I_p)$*;*

(c) If $R_0\beta_*(d) = b_0$ *is true, then the statistic* $S_W(d,h)$ *is asymptotically distributed as* $\chi^2(r)$*, a chi-square distribution with* r *degrees of freedom.*

Theorem 1 shows that the MAGEE procedure has several remarkable features. Theorem 1 (a) ensures that under some conditions detailed in the Appendix, $\hat{\beta}(d,h)$ is a consistent estimate of $\beta_*(d)$ for the adaptive weights in the weighted GEE for any $h \geq 0$. Theorem 1 (b) ensures that $\hat{\beta}(d,h)$ is a $\sqrt{n}$ estimate of $\boldsymbol{\beta}_*(d)$ and asymptotic normal. Theorem 1 (c) ensures that the score test statistic $S_W(d,h_s)$ is asymptotically $\chi^2(r)$ distributed under the null hypothesis $R_0\beta_*(d) = b_0$. These asymptotic properties ensure that it is reliable to apply MAGEE for the analysis of longitudinal imaging data when the sample size is relatively large.

We discuss whether the stochastic adaptive weight defined in (4.12) ensure consistency and asymptotic normality of $\hat{\beta}(d,h)$ at each fixed scale h. To have a better understanding of the MAGEE procedure, we focus on the asymptotic behavior of the adaptive weight as $s = 1$ and then we discuss the scenario with $s > 1$.

Theorem 2. Suppose that Assumptions (C1)-(C5) and (C7) in the Appendix are true. We have the following results for $K_{st}(d,d';n,h)$ *in (4.12):*

(a) $D_\beta(d, d'; h)$ *can be approximated by*

$$D_\beta(d, d'; h_0) = 1(\triangle_*(d, d') = 0) \times O_p(\log(N(\mathcal{D}))) + 1(\triangle_*(d, d') \neq 0) \times n||\triangle_*(d, d') + O_p(\sqrt{\log(N(\mathcal{D}))/n})||_2^2 O_p(1),$$

where $\triangle_*(d, d') = \beta_*(d) - \beta_*(d')$ *and* $N(\mathcal{D})$ *denotes the number of voxels in* $\mathcal{D}$*;*

(b) If $\lim_{n\to\infty} C_n/n = \lim_{n\to\infty} C_n^{-1} \log(N(\mathcal{D})) = \lim_{n\to\infty} C_n^{-1} = 0$*, then we have*

$$\max_{d\in\mathcal{D}} \max_{d'\in B(d,h)\cap\{d':\triangle_*(d,d')\neq 0\}} |K_{st}(d, d'; n, h)| = O_p(\exp(-n)) \to^p 0, \quad \textit{and}$$
$$\max_{d\in\mathcal{D}} \max_{d'\in B(d,h)\cap\{d':\triangle_*(d,d')=0\}} |K_{st}(d, d'; n, h) - 1| \to^p 0.$$

Theorem 2 (a) and (b) show that if the two voxels d and d' have the same true values, then $K_{st}(d, d'; n, h)$ in (4.12) converges to 1. However, if the two voxels d and d' substantially differ from each other, then $K_{st}(d, d'; n, h)$ in (4.12) imposes a decreasing weight on the voxel d'. Thus, $K_{st}(d, d'; n, h)$ in (4.12) can efficiently incorporate information from 'good' voxels, whereas it prevents incorporating information from 'bad' voxels. Particularly, Theorem 2 ensures that assumption (C6) is valid. Thus, the AET procedure with stochastic weights $K_{st}(d, d'; n, h)$ in (4.12) ensures the consistency and asymptotic normality of $\hat{\beta}(d, h)$ at each fixed scale h. Similarly, we can also show that the AET procedure with $K_{st}(d, d'; n, h)$ in (4.18) has the similar property.

In many applications, $\beta(d)$ may be further decomposed as $(\beta_1(d)^T, \beta_2(d)^T)^T$, in which $\beta_1(d)$ is a $q_1 \times 1$ vector of parameters of interest and $\beta_2(d)$ is a $q_2 \times 1$ vector containing additional nuisance parameters. We can calculate $\hat{\beta}(d)$ and then fix $\beta_2(d)$ at $\hat{\beta}_2(d)$ after the initialization step (i). In this way, we only update $\beta_1(d)$ and calculate adaptive weights based on the estimates of $\beta_1(d)$ at each iteration. However, one must modify the weighted GEE method for $\beta_1(d)$ in order to properly account for uncertainty in using $\hat{\beta}_2(d)$, because $\beta_1(d)$ and $\beta_2(d)$ are not 'orthogonal' to each other.

Let $D_i(\beta(d)) = (D_{i,1}(\beta(d)), D_{i,2}(\beta(d)))$, where $D_{i,k}(\beta(d)) = \partial\mu_i(\beta(d))/\partial\beta_k(d)$ for

$k = 1, 2$. We introduce a weighted GEE for $\beta_1(d)$ as follows:

$$\sum_{i=1}^{n} \sum_{d' \in B(d,h)} \omega(d, d'; h) D_{i,1}(\beta_1(d), \hat{\beta}_2(d'))^T V_i((\beta_1(d), \hat{\beta}_2(d)), d')^{-1} e_i(d', \beta_1(d), \hat{\beta}_2(d)) = 0.$$

It will be shown that $\hat{\beta}_2(d)$ does have some effects on the asymptotic distribution of $\hat{\beta}_1(d, h)$. Following the arguments of Theorem 1, we can obtain the asymptotic properties of $\hat{\beta}_1(d, h)$ for $d \in \mathcal{D}$.

Corollary 1. Suppose that Assumptions (C1)-(C7) in the Appendix are true. We have the following results for MAGEE:

(a) $\hat{\beta}_1(d, h)$ converges to $\beta_{*,1}(d)$ in probability;

(b) $\{Cov(\hat{\beta}_1(d, h))\}^{-1/2}[\hat{\beta}_1(d, h) - \beta_*(d)] \to^L N(0, I_{q_1})$, in which $Cov(\hat{\beta}_1(d, h))$ will be given in the Appendix.

Corollary 1 shows that under some mild conditions, the MAGEE ensures that $\hat{\beta}_1(d, h)$ has the desirable asymptotic properties including consistency and asymptotic normality. Thus, MAGEE can smooth solely the image of $\beta_1(d)$, while fixing images of $\beta_2(d)$, $\alpha(d)$ and $\gamma(d)$. For instance, if $\beta_1(d)$ corresponds to diagnosis effect and $\beta_2(d)$ corresponds to age, gender, and other covariates of interest, then MAGEE for only $\beta_1(d)$ allows us to smooth the diagnostic effect image without distorting the images associated with other covariates of interest. This new feature distinguishes our MAGEE significantly from the existing smoothing techniques, which solely smooth the raw images.

4.3 Simulation Studies

We conducted two sets of Monte Carlo simulations to examine the finite sample performance of $\hat{\beta}(d, h)$ and $S_W(d, h)$ with respect to different scales h at the levels of a single voxel and an entire region.

4.3.1 Simulation Studies Part I

We simulated univariate measures across all $m = 4002$ points on the surface of a hippocampus for n subjects. At a given voxel d in $\mathcal{D}$, $y_{ij}(d)$ was simulated according to $y_{ij}(d) = x_{ij}^T\beta(d) + \epsilon_{ij}(d)$ for $j = 1, \cdots, m_i$ and $i = 1, \cdots, n$, where $\beta(d) = (\beta_1(d), \beta_2(d), \beta_3(d))^T$ and $x_{ij} = (1, x_{ij2}, x_{ij3})^T$. We set $m_i = 2$ for $i = 1, \cdots, n/2$ and $m_i = 3$ for $i = n/2 + 1, \cdots, n$. We independently generated $\epsilon_i(d) = (\epsilon_{i1}(d), \cdot, \epsilon_{im_i}(d))^T$ from a multivariate $N(0, \Omega)$ distribution, where diag(Ω) equals a $m_i \times 1$ vector with all ones and the correlation between $\epsilon_{ij_1}(d)$ and $\epsilon_{ij_2}(d)$ equals $0.7^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \cdots, m_i$ and $i = 1, \cdots, n$. We generated x_{i12}, x_{i22}, and x_{i32} from $U[0, 1]$, $U[1, 2]$, and $U[2, 3]$, respectively, where $U[a, b]$ denotes the uniform distribution on [a, b]. x_{ij3} was a time invariant covariate representing diagnostic effect generated independently from a Bernoulli distribution with equal probability for each i. We also created ROI1 and ROI2, which are two nested circles with radius at 3 and 5, respectively, and labeled the region outside of ROI1 and ROI2 as ROI3. We set $(\beta_1(d), \beta_3(d)) = (1, 1)$ across all voxels, whereas we set $\beta_2(d)$ as 0 in ROI3, 1 in ROI2, and 2 in ROI3, respectively (Fig. 4.2). We chose two sample sizes: $n = 50$ and $n = 80$.

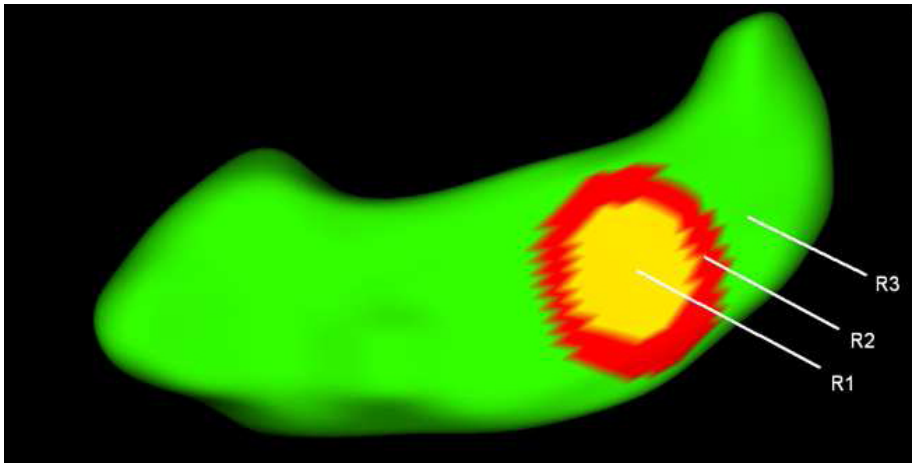

Figure 4.1: Simulation study parts I: three regions of interest ($R1$: ROI1 with yellow color; $R2$: ROI2 with red color; $R3$: ROI3 with green color) on a reference hippocampus.

We fitted GEE with $E[y_{ij}(d)] = x_{ij}^T\boldsymbol{\beta}(d)$ and AR(1) working correlation structure.

Table 4.1: Bias ($\times 10^{-3}$), RMS($\times 10^{-1}$), SD($\times 10^{-1}$), and RS of $\boldsymbol{\beta}$ parameters. BIAS denotes the bias of the mean of the MARM estimates; RMS denotes the root-mean-square error; SD denotes the mean of the standard deviation estimates; RS denotes the ratio of RMS over SD.

	β_1				β_2				β_3			
	BIAS	RMS	SD	RS	BIAS	RMS	SD	RS	BIAS	RMS	SD	RS
						$n = 50$						
ROI1: h_0	0.11	2.62	2.61	1.01	1.30	0.85	0.83	1.02	-4.86	2.55	2.48	1.03
ROI1: h_5	0.54	0.77	0.75	1.03	0.93	0.25	0.24	1.03	-3.54	0.75	0.71	1.05
ROI2: h_0	-2.00	2.64	2.61	1.01	0.77	0.84	0.83	1.01	2.78	2.55	2.48	1.03
ROI2: h_5	-1.09	0.72	0.70	1.03	0.84	0.23	0.22	1.02	0.90	0.71	0.66	1.07
ROI3: h_0	0.09	2.66	2.60	1.02	0.09	0.84	0.83	1.02	0.15	2.54	2.48	1.03
ROI3: h_5	0.11	0.74	0.71	1.05	0.08	0.23	0.23	1.04	0.20	0.72	0.67	1.07
						$n = 80$						
ROI1: h_0	1.15	2.11	2.09	1.01	-0.72	0.62	0.63	0.98	1.53	2.08	2.00	1.04
ROI1: h_5	-0.07	0.70	0.69	1.01	5.30	0.24	0.21	1.14	2.73	0.68	0.66	1.03
ROI2: h_0	2.83	2.10	2.09	1.00	-0.57	0.63	0.63	1.00	-0.37	2.02	2.00	1.01
ROI2: h_5	2.96	0.68	0.67	1.01	- 2.62	0.21	0.20	1.03	-0.62	0.65	0.63	1.03
ROI3: h_0	-0.09	2.11	2.09	1.01	0.01	0.64	0.63	1.01	0.34	2.03	2.00	1.01
ROI3: h_5	-0.03	0.59	0.57	1.03	0.03	0.18	0.17	1.03	0.20	0.57	0.55	1.04

We used MAGEE to adaptively calculate the parameter estimates across all voxels at 6 different scales. Our primary interest is to make inference on $\beta(d)$ and other parameters such as $\alpha(d)$ in AR(1) are regarded nuisance parameters and fixed at their estimators after the initialization step. In each ROI, we calculated the bias, the empirical standard errors (RMS), and the mean of the standard error estimates (SD) based on the results from the 1,000 simulated hippocampus data sets. We observed the following results. The biases are similar at h_0 and h_5. The RMS and SD at h_5 are much smaller than those at h_0. In addition, the RMS and its corresponding SD are relatively close to each other at both the h_0 and h_5 scales in each of the three ROIs (Table 4.1). As expected, increasing n decreases the RMS and SD of the parameter estimates.

4.3.2 Simulation Studies Part II

Following the setup in Section 4.3.1, we simulated an additional dataset at all the $m = 4002$ points on the surface of a hippocampus for 50 subjects, except that six new ROIs were constructed as three sets of nested circles. In the first set of nested circles, $\beta_2(d)$ were set at 1 and 2 in the inner and outer circles with radii being 2 and 4, respectively. In the second set of nested circles, $\beta_2(d)$ were set at 0.6 and 0.8 in the inner and outer circles at radii being 3 and 5, respectively. In the third set of nested circles, $\beta_2(d)$ was set to 0.4 and 0.6 in the inner and outer circles with radii being 3 and 4, respectively.

The parameter $\beta_2(d)$ outside of these six ROIs was always set at 0. We used MAGEE to calculate the estimate of β at 6 different scales. It is clear that the estimate of $\beta_2(d)$ is more precisely estimated at h_5 compared with at h_0 at different signal to noise ratios (Fig. 4.2 (a) and (c)). Similarly, the p-value map generated for testing $H_0 : \beta_2(d)$=0 at h_5 also performs much better than that at $h_0 = 0$ (Fig. 4.2(b) and (d)).

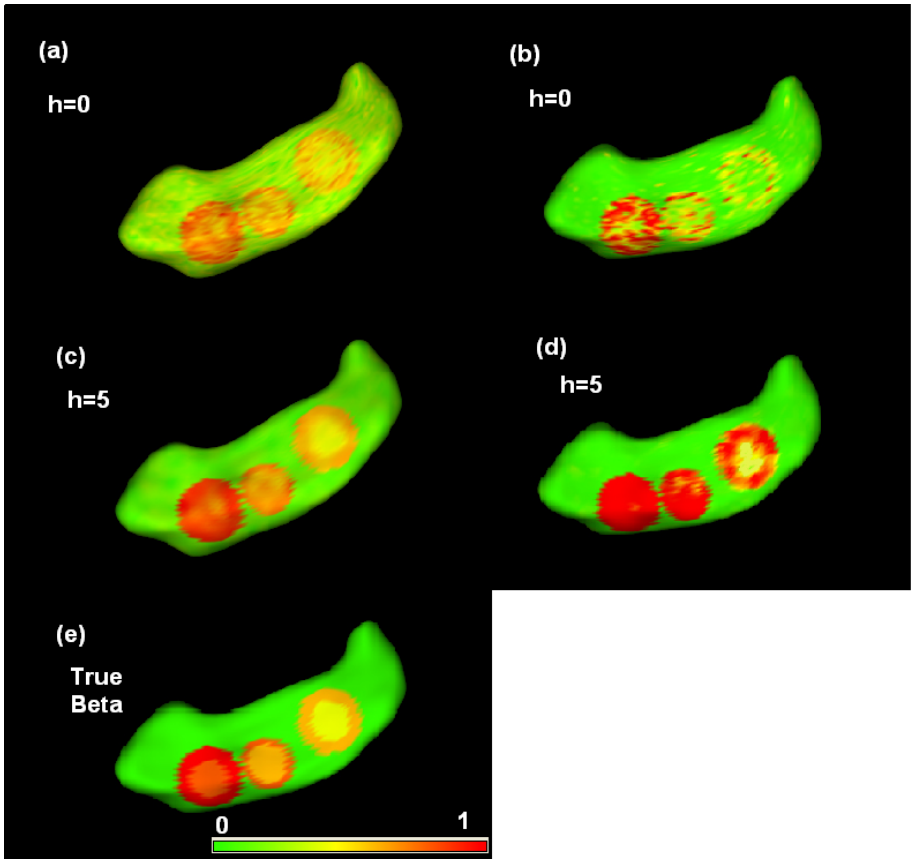

Figure 4.2: Comparison of the voxel-wise approach and MAGEE for the simulated hippocampus dataset with three sets of nested circles (panel (e)): the maps of resampling corrected $-\log_{10}(p)$ values and estimated parameters $\beta_2(d)$ based on the voxel-wise GEE approach (panels (a) and (b)) and MAGEE (panels (c) and (d)).

4.4 Real Data Analysis

A wealth of cross-sectional diffusion tensor imaging (DTI) studies has been conducted on characterizing white matter development (prenatal to adolescent stages) using various

DTI parameters such as fractional anistropy (FA) and radial (RD) diffusivity in the past decade. Current DTI studies including neonates have revealed three phases in the early postnatal brain development: the rapid changes within first 12 months, the slow maturation from 12 to 24 months, and the steady state afterwardss. Particularly, in white matter, neonates have significantly lower anistropy values and significantly higher MD values compared to adults (Neil et al., 1998; Zhai et al., 2003). These DTI studies also reveal the temporal non-linearity and spatial inhomogeneity of the apparent changes in DTI parameters within brain (Mukherjee et al., 2001; Mukherjee et al., 2002; Schneider et al., 2004).

We used 38 subjects from a larger study designated to the investigation of early brain development led by Dr. Gilmore at the University of North Carolina at Chapel Hill. For each subject, diffusion-weighted images were acquired at 2 weeks, year 1 and year 2. Diffusion gradients with a $b-$value of 1000 s/mm^2 were applied in six non-collinear directions, (1,0,1), (-1,0,1), (0,1,1), (0,1,-1), (1,1,0), and (-1,1,0) and a $b = 0$ reference scan. Forty-six contiguous slices with a slice thickness of 2 mm covered a field of view (FOV) of 256×256 mm^2 with an isotropic voxel size of $2 \times 2 \times 2$ mm^3. A total of eighteen acquisitions were used to improve the signal-to-noise ratio (SNR). High resolution T1 weighted (T1W) images were acquired using a 3D MP-RAGE sequence.

We then calculated a weighted least squares estimation method to construct the diffusion tensors (Basser, Mattiello, and LeBihan 1994 b; Zhu *et al.*, 2007b). All images were visually inspected before analysis to ensure no bulk motion. All DT images (38 subjects, 3 time points each) were registered, using TIMER, onto a randomly selected brain DT image of a 2-year-old subject. The aligned images were then voxel-wise averaged to create the mean DT image, from which the FA map can be computed (Yap *et al.*, 2009).

Fractional anisotropy (FA) calculated from DTIs has been 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 use FA images to characterize the spatial pattern of white matter maturation. We fitted GEE with $E[y_{ij}(d)] = \beta_0(d)+t_i\beta_1(d)$ and the AR(1) working correlation structure at each voxel of the template. We applied the MAGEE procedure with $c_h = 1.25$ and $S = 6$ to carry out the statistical analysis and tested $H_0 : \beta_1(d) = 0$ for time effect across all voxels d. We treated other parameters (e.g., the parameter in the AR(1)) as nuisance parameters and fixed them after the initialization iteration. Compared with the results from the standard voxel-wise method at h_0 (Fig. 4.3 (a)-(d)), MAGEE shows a clear advantage in detecting more significant and smooth activation areas as the bandwidth h increases (Fig. 4.3 (e)-(h)).

To identify different spatial patterns of white matter maturation, we further clustered the growth trajectories according to the two dimensional features $(\beta_0(d), \beta_1(d))$ across the template. Standard mixture package from SPM8 was used to cluster the two-dimensional data and to choose 5 as the optimal number of clusters (Fig. 4.4 (a)). These 5 clusters well represent the gray matter, the boundary of gray matter and white matter and 3 components of white matter. To show the superiority of MAGEE, the clustering results based on the MAGEE estimates from the scale 5 are visually more smoother than the clustering results based on the MAGEE estimates from scale 0 (Fig 4.4 (b)). We also compared the probability maps for each of these 5 clusters at scale 5 and scale 0. The probability maps also show more smooth pattern for scale 5 versus scale 0 (Fig. 4.5).

4.5 Discussion

This article has developed a unified estimation and smoothing procedure for the spatial and adaptive analysis of neuroimaging data from longitudinal studies. We have demonstrated its superiority over the voxel-wise approach using simulated and real imaging data. MAGEE is semiparametric, spatial, hierarchical and adaptive. MAGEE can adaptively utilize available information in the neighboring voxels to increase the precision of parameter estimates and the power of test statistics in detecting subtle changes

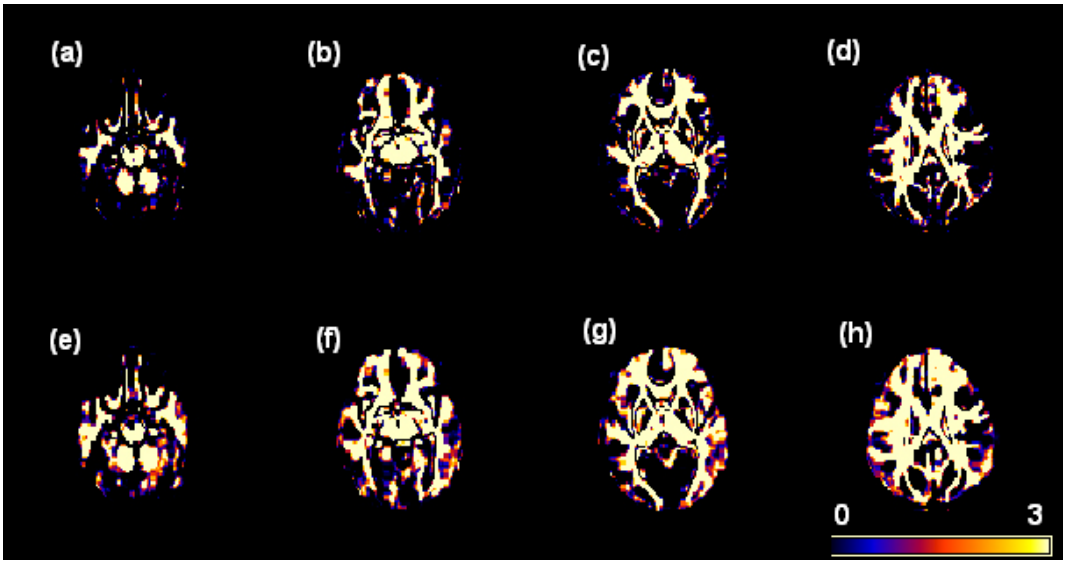

Figure 4.3: Results from the neonatal project on brain development. Panels (a), (b),(c) and (d) : the corrected $-\log_{10}(p)$ values of the Score test statistics $S_W(d, h_0)$ from three selected slices; panels (e), (f),(g) and (h): the corrected $-\log_{10}(p)$ values of the Score test statistics $S_W(d, h_5)$ from the selected slices; (I) the comparison of the histograms for $S_W(d, h_0)$ and $S_W(d, h_5)$ across all voxels.

of brain structure and function. We have shown that MAGEE can adaptively smooth images while accounting for the spatial pattern of activation regions. We have shown that MAGEE can simultaneously smooth all raw images from all subjects across all time points using the learned information during the statistical estimation step, while MAGEE can smooth images of all parameters of interest after fixing images of other nuisance parameters. We have theoretically examined the adaptive weights in the MAGEE and formally establish the asymptotic properties including consistency and asymptotic distributions of the parameter estimates and test statistics for MAGEE.

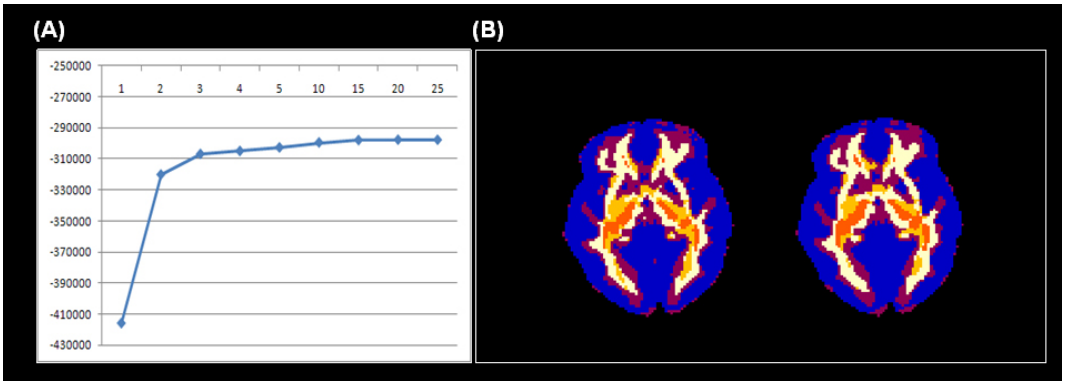

Figure 4.4: Clustering results for the neonatal project on brain development. Panel (a): 5 clusters are the optimal clusters selected by negative free energy criteria. Panel (b): Clustering maps show 5 components for scale at 0 (left) and scale at 5 (right).

4.6 Appendix

We need some notation. To avoid notational complexity, we assume that all nuisance parameters $(\alpha(d), \gamma(d))$ are known. Let $V_i(\beta, d')$ denote $V_i(\alpha(d'), \beta, \gamma(d'))$. We define

$$\begin{aligned} M_n(\beta(d)) &= n^{-1}\sum_{i=1}^{n} D_i(\beta(d))^T V_i^{-1}(\beta(d), d) D_i(\beta(d)), \\ H_n(\beta(d)) &= n^{-1}\sum_{i=1}^{n} D_i(\beta(d))^T V_i^{-1}(\beta(d), d)\Sigma_i(\beta(d)) V_i(\beta(d), d)^{-1} D_i(\beta(d)), \quad \text{and} \\ g_{n,i}(\beta(d), d') &= D_i(\beta(d))^T V_i^{-1}(\beta(d), d')[Y_i(d') - \mu_i(\beta(d))]. \end{aligned}$$

We also define $\mathcal{D}_*(d,h)^c = \{d' \in B(d,h) : \triangle_*(d,d') \neq 0\}$ and $\mathcal{D}_*(d,h) = \{d' \in B(d,h) : \triangle_*(d,d') = 0\}$, where $\triangle_*(d,d') = \beta_*(d) - \beta_*(d')$.

The following assumptions are needed to facilitate development of our methods, although they are not the weakest possible conditions.

(C1) $1 \geq \omega(d,d';h) \geq 0$ and $\omega(d,d;h) = 1$ for all $d, d' \in \mathcal{D}$ and $h \geq 0$.

(C2) Let $Z_i = (x_{i1}, \cdots, x_{im_i}, Y_{i,\mathcal{D}})\}$ for $i = 1, \cdots, n$. The data $Z_1, \cdots, Z_n$ form independent clusters.

(C3) For any $d \in \mathcal{D}$, there is an unique interior point of $\mathcal{B}$, denoted by $\beta_*(d)$, such that $E[Y_i(d)|x_{i1}, \cdots, x_{im_i}] = \mu_i(\beta_*(d))$ for all i, where $\mathcal{B}$ is a compact set in R^q and

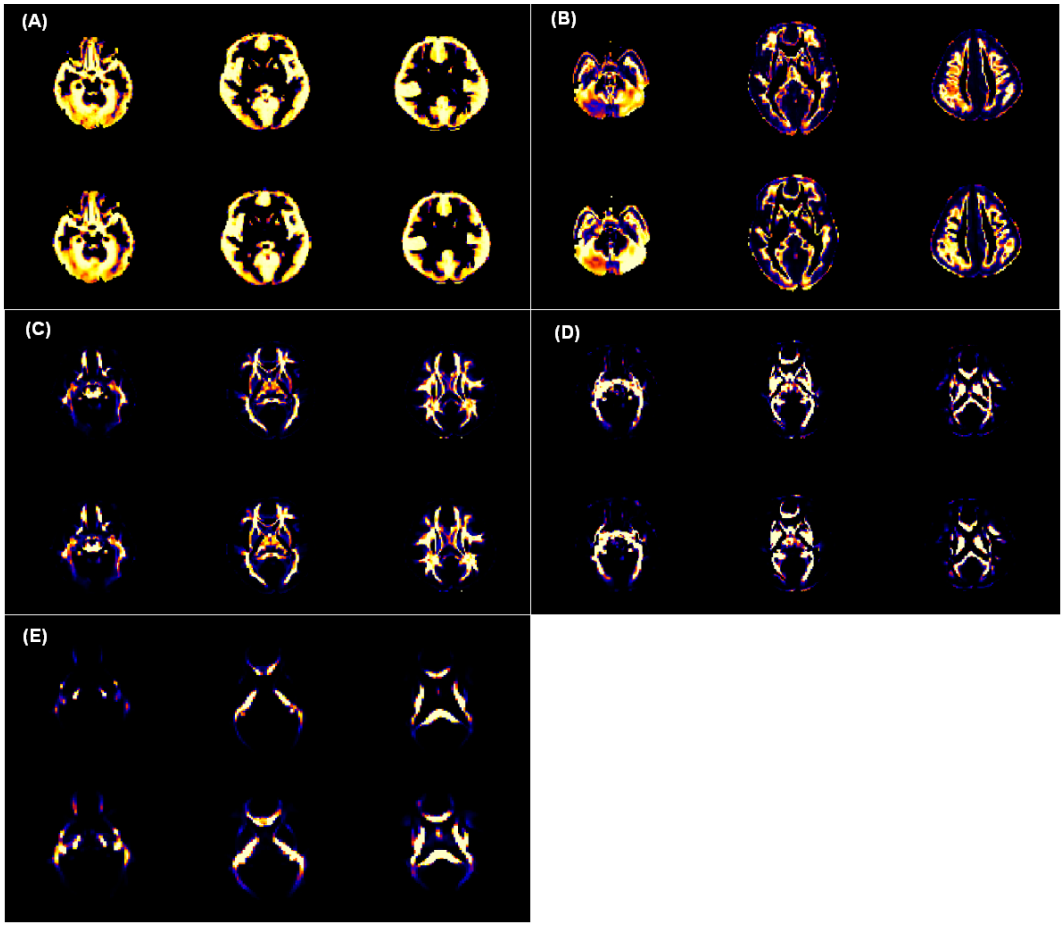

Figure 4.5: Probability maps for five clusters for the neonatal project on brain development. The upper row of each panel of (A)-(E) shows the three selected probability maps based on the results obtained from MAGEE at scale 0, whereas the lower row of each panel of (A)-(E) presents the three selected probability maps based on the results obtained from MAGEE at scale 5.

the expectation is taken with respect to the true conditional distribution of $Y(d)$ given covariate.

(C4) For all voxels $d \in \mathcal{D}$, $g_{n,i}(\beta(d), d)$ is continuously differentiable on $\mathcal{B}$. For all $j, k, l = 1, \cdots, p$, $||g_{n,i}(\beta(d), d)||_2^2$, and $||\partial_{\beta(d)} g_{n,i}(\beta(d), d)||_2$ are dominated by an integral function $g(Y(d), x)$ such that $E[\max_{d\in\mathcal{D}} |g(Y(d), x)|^2] < \infty$ and

$$E[\max_{d\in\mathcal{D}} |g(Y(d), x)|^2 1(\max_{d\in\mathcal{D}} |g(Y(d), x)| > \eta\sqrt{n})] \to 0$$

for every $\eta > 0$.

(C5) For all $d \in \mathcal{D}$, $\lim_{n\to\infty} M_n(\beta_*(d)) = M(\beta_*(d))$ and $\lim_{n\to\infty} H_n(\beta_*(d)) = H(\beta_*(d))$. For a fixed $\delta > 0$,

$$\begin{aligned}
\infty &> \sup_{d\in\mathcal{D}} \max_{\beta(d)\in B(\beta_*(d),\delta)} (\lambda_{\max} M_n(\beta(d))) \geq \inf_{d\in\mathcal{D}} \min_{\beta(d)\in B(\beta_*(d),\delta)} (\lambda_{\min} M_n(\beta(d))) > 0, \\
\infty &> \sup_{d\in\mathcal{D}} \max_{\beta(d)\in B(\beta_*(d),\delta)} (\lambda_{\max} H_n(\beta(d))) \geq \inf_{d\in\mathcal{D}} \min_{\beta(d)\in B(\beta_*(d),\delta)} (\lambda_{\min} H_n(\beta(d))) > 0,
\end{aligned}$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix, respectively.

(C6) Let $\max_{d\in\mathcal{D}} \max_{d'\in\mathcal{D}_*(d,h)^c} K_{st}(d,d';n,h) = K_{st}^c(n,h)$. As $n \to \infty$,

$$\sqrt{n} K_{st}^c(n,h) \to^p 0 \text{ and } \max_{d\in\mathcal{D}} \max_{d'\in\mathcal{D}_*(d,h)} |K_{st}(d,d';n,h) - K_{st}(d,d',h)| \to^p 0,$$

where $\to^p$ denotes convergence in probability and $K_{st}(d,d',h)$ is a nonrandom function of (d,d',h).

(C7) $\lim_{n\to\infty} \log(N(\mathcal{D}))/n = 0$.

Remarks A1: For each fixed $d \in \mathcal{D}$, Assumptions (C3)-(C5) are generalizations of the standard conditions for ensuring first order asymptotic properties (e.g., consistency and asymptotic normality) of Z-estimators (van der Vaart, 1998). Assumption (C3) is an identification condition, whereas Assumption (C4) is a uniform smoothness and integration condition. Particularly, Assumption (C4) ensures that $|g_{n,i}(\beta(d),d)|^2$ and $|\partial_{\beta(d)} g_{n,i}(\beta(d),d)|$ are uniformly integrable for all $d \in \mathcal{D}$. Assumption (C5) is needed to ensure that the covariance matrix of $\hat{\beta}(d,h)$ is positive definite for all $d \in \mathcal{D}$.

Remarks A2: Assumption (C6) is needed just for ensuring desirable asymptotic properties of $\hat{\beta}(d,h)$ based on the weighted GEE. Assumption (C6) requires that for two similar voxels d and d', $K_{st}(d,d;n,h)$ converge to a nonrandom $K_{st}(d,d',h)$, whereas for two distinctive voxels, $K_{st}^c(n,h)$ converge to zero faster than $\sqrt{n}$. We have already proposed two different choices of $K_{st}(d,d;n,h)$ and will examine whether they satisfy Assumption (C6). Actually, we can show that for $K_{st}(d,d',h)$ in (4.12) satisfies Assumption (C6)

such that $K_{st}(d, d', h) = 1$.

Remarks A3: In neuroimaging data, although $N(\mathcal{D})$ is much larger than the sample size n, Assumption (C7) claims that we just need a relatively large sample size compared to $\log(N(\mathcal{D}))$. For instance, in most neuroimaging data, $N(\mathcal{D}) \approx 100^3$ and $\log(10^3) = 14$. Therefore, a sample size such as 100 may be reasonable to use asymptotic normality in making statistical inferences for MAGEE. Assumption (C7) is needed to invoke maximal inequalities (var der Vaart and Wellner, 1996).

Derivation of Score Test Statistic. To consider the test statistic $S_W(d, h)$, we need additional notation as follows:

$$\begin{aligned} G_n(\beta(d); \omega, h) &= \sum_{d' \in B(d,h)} \omega(d, d'; h) \sum_{i=1}^{n} D_i(\beta(d))^T V_i^{-1}(\beta(d), d')[Y_i(d') - \mu_i(\beta(d))], \\ \partial_\beta G_n(\beta(d); \omega, h) &\approx -\sum_{d' \in B(d,h)} \omega(d, d'; h) \sum_{i=1}^{n} D_i(\beta(d))^T V_i^{-1}(\beta(d), d') D_i(\beta(d)). \end{aligned}$$

Without loss of generality, we assume that $R = (R_1, R_2)$, in which R_1 is an $r \times r$ nonsingular matrix and R_2 is an $r \times (q - r)$ matrix. Let $\beta = (\beta_{(1)}^T, \beta_{(2)}^T)^T$, where $\beta_{(1)}$ is an $r \times 1$ vector corresponding to R_1 and $\beta_{(2)}$ is a $(q-r) \times 1$ vector corresponding to R_2. If we define

$$\nu_1 = R_1\beta_{(1)} + R_2\beta_{(2)} - b_0 \text{ and } \nu_2 = (\beta_{(2)}), \tag{4.20}$$

then there exists a one-to-one correspondence between $(\nu_1, \nu_2) = f(\beta)$ and $\beta = f^{-1}(\nu_1, \nu_2)$. Thus, we have

$$\frac{\partial(\beta_{(1)}, \beta_{(2)})}{\partial(\nu_1, \beta_{(2)})} = \begin{pmatrix} R_1^{-1} & -R_1^{-1}R_2 \\ 0 & I_{q-r} \end{pmatrix}.$$

Let $\partial_{\nu(d)} = \partial/\partial_{\nu(d)}$. We define

$$\begin{aligned} G_n(\nu(d);\omega,h) &= \begin{pmatrix} G_{n,1}(\nu(d);\omega,h) \\ G_{n,2}(\nu(d);\omega,h) \end{pmatrix} \\ &= \sum_{d'\in B(d,h)} \omega(d,d';h) \sum_{i=1}^{n} \partial_{\nu(d)}\mu_i(\beta(d))^T V_i^{-1}(\beta(d),d') e_i(d',\beta(d)), \\ -\partial_{\nu(d)} G_n(\nu(d);\omega,h) &\approx \Sigma(\nu_1,\nu_2) = \begin{pmatrix} \Sigma_{\nu_1\nu_1} & \Sigma_{\nu_1\nu_2} \\ \Sigma_{\nu_2\nu_1} & \Sigma_{\nu_2\nu_2} \end{pmatrix} \\ &\approx \sum_{d'\in B(d,h)} \omega(d,d';h) \sum_{i=1}^{n} \partial_{\nu(d)}\mu_i(\beta(d))^T V_i^{-1}(\beta(d),d') \partial_{\nu(d)}\mu_i(\beta(d)), \end{aligned}$$

where $G_{n,1}(\nu(d);\omega,h)$ and $G_{n,2}(\nu(d);\omega,h)$ are the $r\times 1$ and $(q-r)\times 1$ subcomponents of $G_n(\nu(d);\omega,h)$ corresponding to $\nu_1(d)$ and $\nu_2(d)$, respectively.

Let $\nu_*(d) = (0,\nu_{2*}(d))$ be the true parameter vector of $\beta(d)$ under $H_0(d)$ and $\tilde{\nu}(d) = (0,\tilde{\nu}_2(d))$ is the maximum quasi-likelihood estimate of $\nu(d)$ under $H_0(d)$. We use a Taylor's series expansion to obtain

$$0 = G_{n,2}(\tilde{\nu}(d);\omega,h) \approx G_{n,2}(\nu_*(d);\omega,h) + \partial_{\nu_2(d)} G_{n,2}(\nu_*(d);\omega,h)[\tilde{\nu}_2(d) - \nu_{2*}(d)].$$

Thus, we have $\tilde{\nu}_2(d) - \nu_{2*}(d) \approx \Sigma_{\nu_2\nu_2}^{-1} G_{n,2}(\nu_*(d);\omega,h)$. We apply a Taylor's expansion to obtain

$$G_{n,1}(\tilde{\nu}(d);\omega,h) \approx G_{n,1}(\nu_*(d);\omega,h) - \Sigma_{\nu_1\nu_2}\Sigma_{\nu_2\nu_2}^{-1} G_{n,2}(\nu_*(d);\omega,h) \approx \sum_{i=1}^{n} \hat{U}_{i,\omega}(\tilde{\beta}(d,h)),$$

where $\hat{U}_{i,\omega}(\tilde{\beta}(d,h))$ is given by

$$[I_r \ \vdots \ -\Sigma_{\nu_1\nu_2}\Sigma_{\nu_2\nu_2}^{-1}] \partial_{\nu(d)}\mu_i(\beta(d))^T V_i^{-1}(\beta(d),d') e_i(d',\beta(d)). \tag{4.21}$$

Proof of Theorem 1. We prove Theorem 1 (a) and (b) by induction. The proof primarily

consists of three major steps: (i) $s = 0$; (ii) $s = 1$; (iii) $s \geq 1$.

The proof of Step (i) for $s = 0$ consists of two steps. In Step (1.1), we will show that $\hat{\beta} = (\hat{\beta}(d) : d \in \mathcal{D})$ converges $\beta_* = (\beta_*(d) : d \in \mathcal{D})$ in probability. We need to introduce some notation. Let T be a bounded brain region in R^g containing all voxels $d \in \mathcal{D}$, where $g = 2$ for the 2D surface and $g = 3$ for the 3D volume. Let $\mathcal{B}^{\mathcal{D}} = \prod_{d \in \mathcal{D}} \mathcal{B}$ be the parameter space for β and $\ell^\infty(T)^q$ is the product of q $\ell^\infty(T) = \{z : T \to R, \quad \sup_{t \in T} |z(t)| < \infty\}$. Let $\Psi_n : \mathcal{B}^{\mathcal{D}} \to \ell^\infty(T)^q$ and $\Psi : \mathcal{B}^{\mathcal{D}} \to \ell^\infty(T)^q$ be random maps and a deterministic map, respectively, such that

$$\Psi_n(\beta)(t) = n^{-1} \sum_{i=1}^{n} g_{n,i}(\beta(d_t)) \quad \text{and} \quad \Psi(\beta)(t) = E[g_{n,i}(\beta(d_t))],$$

in which d_t denotes the voxel covering t.

To prove the consistency of $\hat{\beta}$, we will show that

$$\begin{aligned} &\sup_{\beta \in \mathcal{B}^{\mathcal{D}}} \sup_{t \in T} ||\Psi_n(\beta)(t) - \Psi(\beta)(t)||_2 \to 0 \quad \text{and} \\ &\inf_{\beta \in \mathcal{B}^{\mathcal{D}} : ||\beta - \beta_*|| \geq \epsilon} \sup_{t \in T} ||\Psi(\beta)(t)||_2 > \sup_{t \in T} ||\Psi(\beta_*)(t)||_2. \end{aligned} \tag{4.22}$$

It follows from Assumptions (C3) and (C4) that the second term in equation (4.22) is true. To prove the first term in equation (4.22), we note that

$$\sup_{\beta \in \mathcal{B}^{\mathcal{D}}} \sup_{t \in T} ||\Psi_n(\beta)(t) - \Psi(\beta)(t)||_2 = \max_{d \in \mathcal{D}} A_n(d), \tag{4.23}$$

where $A_n(d) = \sup_{\beta(d) \in \mathcal{B}} |n^{-1} \sum_{i=1}^{n} \{g_{n,i}(\beta(d)) - E[g_{n,i}(\beta(d))]\}|$. Then, we consider $\mathcal{F} = \{g_{n,i}(\beta(d)) : d \in \mathcal{D}, \beta(d) \in \mathcal{B}\}$ with an envelope $\max_{d \in \mathcal{D}} g(Y(d), x)$. Following the arguments in Theorem 2.4.3 of van der Vaart and Wellner (1996), we can show that

$E[\max_{d\in\mathcal{D}} A_n(d)]$ is bounded above by

$$\sqrt{[1+p\log(C_1(\epsilon)K)+\log(N(\mathcal{D}))]/n}C_2K+$$
$$2E[\max_{d\in\mathcal{D}} g(Y(d),x)1\{\max_{d\in\mathcal{D}} g(Y(d),x)>K\}]+\epsilon\to 0,$$

where C_2 is a constant independent of ϵ, K can be chosen such that the second term of the above equation is arbitrarily small, and $C_1(\epsilon)$ is a constant depending on ϵ. Finally, following the arguments in Theorems 5.7 and 5.9 of van der Vaart (1998), we can prove consistency of $\hat{\beta}$.

In Step (1.2), we will prove the asymptotic normality of $\sqrt{n}(\hat{\beta}-\beta_*)$. For each $d\in\mathcal{D}$, a Taylor's series expansion gives

$$0=\Psi_n(\hat{\beta})(d)=\Psi_n(\beta_*)(d)+\partial_{\beta(d)}\Psi_n(\tilde{\beta})(d)[\hat{\beta}(d,h_0)-\beta_*(d)], \tag{4.24}$$

where $\tilde{\beta}\in\mathcal{B}^{\mathcal{D}}$ and $\tilde{\beta}(d)$ is on the line connecting $\beta(d)$ and $\beta_*(d)$. Similar to the proof of (4.23), we can show that

$$\sup_{\beta\in\mathcal{B}^{\mathcal{D}}:||\beta-\beta_*||_2\le\epsilon}\sup_{t\in T}||\partial_{\beta(d_t)}\Psi_n(\beta)(t)-\partial_{\beta(d_t)}\Psi(\beta)(t)||_2\to 0 \tag{4.25}$$

in probability, when $\log(N(\mathcal{D}))/n$ is sufficiently small. Therefore, we can show that

$$\sqrt{n}[\hat{\beta}(d,h_0)-\beta_*(d)]=[-\partial_{\beta(d)}\Psi(\beta_*)(d)+o_{p,\mathcal{D}}(1)]^{-1}\sqrt{n}\Psi_n(\beta_*)(d), \tag{4.26}$$

for all $d\in\mathcal{D}$, where $o_{p,\mathcal{D}}(1)$ denotes uniform convergence to zero for all $d\in\mathcal{D}$. It is easy to prove the asymptotic normality of $\sqrt{n}[\hat{\beta}(d,h_0)-\beta_*(d)]$ for each $d\in\mathcal{D}$. Furthermore, by using Theorem 2.14.1 of van der Vaart and Wellner (1996), we can show that $\sup_{d\in\mathcal{D}}||\Psi_n(\beta_*)(d)||_2=O_p(\sqrt{\log(N(\mathcal{D}))/n})$, which yields

$$\max_{d\in\mathcal{D}}||\hat{\beta}(d,h_0)-\beta_*(d)||_2=O_p(\sqrt{\log N(\mathcal{D})/n}). \tag{4.27}$$

We prove Step (ii) for $s = 1$ as follows. Let $\tilde{\omega}(d, d'; h_1) = \omega(d, d'; h_1) / \sum_{d' \in \mathcal{B}} \omega(d, d'; h_1)$. It follows that

$$\sup_{\beta(d) \in \mathcal{B}} |n^{-1} G_n(\beta(d); \tilde{\omega}, h_1) - G(\beta(d); \tilde{\omega}, h_1)| \leq \sum_{d' \in B(d, h_1)} \tilde{\omega}(d, d'; h_1) \delta_n(d') \leq \max_{d' \in B(d, h_1)} \delta_n(d'),$$

where $\delta_n(d) = \sup_{\beta(d) \in \mathcal{B}} |n^{-1} \sum_{i=1}^n g_{n,i}(\beta(d)) - n^{-1} E[\sum_{i=1}^n g_{n,i}(\beta(d))]|$. Then, following arguments in Theorems 2.7.11 and 2.4.3 of van der Vaart and Wellner (1996) and assumptions (C2)-(C4), we can show that

$$\begin{aligned} & E[\max_{d \in \mathcal{D}} \delta_n(d)] \leq \sqrt{[1 + p \log(C_1(\epsilon) K) + \log(N(\mathcal{D}))]/n} C_2 K \\ & + 2E[\max_{d \in \mathcal{D}} g(Y(d), x) 1\{\max_{d \in \mathcal{D}} g(Y(d), x) > K\}] + \epsilon \to 0. \end{aligned}$$

Since the above arguments are independent of $\tilde{\omega}(d, d'; h_1)$, we can conclude that

$$\max_{d \in \mathcal{D}} \sup_{\beta(d) \in \mathcal{B}} |n^{-1} G_n(\beta(d); \tilde{\omega}, h_1) - G(\boldsymbol{\theta}(d); \tilde{\omega}, h_1)| \to 0 \tag{4.28}$$

in probability, and it holds for any adaptive weights $\tilde{\omega}(d, d'; h)$.

It follows from (4.27) that

$$\begin{aligned} & \max_{d \in \mathcal{D}} \sup_{\beta(d)} |n^{-1} G(\beta(d); \tilde{\omega}, h_1) - \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h) \sum_{i=1}^n E[g_{n,i}(\beta(d'))]| \\ \leq \; & K_{st}^c(n, h) E[\max_{d \in \mathcal{D}} g(Y(d), x)] \to 0. \end{aligned} \tag{4.29}$$

Since $0 = \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h) \sum_{i=1}^n E[g_{n,i}(\beta_*(d))]$, it follows from Theorem 5.7 of var der Vaart (1998) and the arguments in the proof of Theorem 1 (a) that $\hat{\beta}(h_1) = (\hat{\beta}(d, h_1) : d \in \mathcal{D})$ converges to β_* in probability.

To prove asymptotic normality of $\hat{\beta}(d, h_1)$, we can use a Taylor's series expansion to

show that

$$0 = G_n(\hat{\beta}(d,h_1);\tilde{\omega},h_1) = G_n(\hat{\beta}_*(d);\tilde{\omega},h_1) + \partial_{\beta(d)}G_n(\tilde{\beta}_*(d);\tilde{\omega},h_1)[\hat{\beta}(d,h_1) - \beta_*(d)],$$

where $\tilde{\beta}(d,h_1)$ is on the segment joining $\hat{\beta}(d,h_1)$ and $\beta_*(d)$. Similar to the arguments in the proof of Theorem 1 (b) and (4.29), we can show that

$$\begin{aligned}
&\max_{d\in\mathcal{D}} \sup_{||\beta_*(d)-\beta(d)||_2\le\epsilon} |n^{-1}\partial_{\beta(d)}G_n(\beta(d);\tilde{\omega},h_1) - \\
&\quad \sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \tilde{\omega}(d,d';h_1)n^{-1}\sum_{i=1}^n E[\partial_{\beta(d)}g_{n,i}(\beta(d),d')]| \to 0, \\
&\max_{d\in\mathcal{D}} n^{-1/2}|G_n(\beta_*(d);\tilde{\omega},h_1) - \sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \tilde{\omega}(d,d';h_1)\sum_{i=1}^n g_{n,i}(\beta_*(d),d')| \\
\le\ & n^{1/2}K_{st}^c(n,h)E[\sup_{d\in\mathcal{D}} g(Y(d),x)]O(1) \to 0.
\end{aligned}$$

Finally, $\sqrt{n}[\hat{\beta}(d,h_1) - \beta_*(d)]$ can be represented as

$$\begin{aligned}
&\{-\sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \tilde{\omega}(d,d';h_1)n^{-1}\sum_{i=1}^n E[\partial_{\beta(d)}g_{n,i}(\beta_*(d),d')] + o_{p,\mathcal{D}}(1)\}^{-1} \times \\
&n^{-1/2}\sum_{d'\in B(d,h_1)\cap\mathcal{D}_*(d)} \tilde{\omega}(d,d';h_1)\sum_{i=1}^n g_{n,i}(\beta_*(d),d'). \qquad (4.30)
\end{aligned}$$

By using Theorem 2.14.1 of van der Vaart and Wellner (1996), we can show that

$$\max_{d\in\mathcal{D}} ||n^{-1/2}\sum_{i=1}^n g_{n,i}(\beta_*(d),d')||_2 = O_p(\sqrt{\log N(\mathcal{D})}),$$

which yields that $\max_{d\in\mathcal{D}} ||\hat{\beta}(d,h) - \beta_*(d)||_2 = O_p(\sqrt{\log N(\mathcal{D})/n})$.

In Step (iii), by using induction and the above arguments in Step (ii), we can prove Theorem 1 (a) and (b) for any fixed $s > 1$.

Given the results in Theorem 1 (a) and (b), we can apply standard arguments in the literature to prove Theorem 1 (c). We omit the details for simplicity.

Proof of Theorem 2. Let $\hat{\Delta}(d,h) = \hat{\beta}(d,h) - \beta_*(d)$. Since $D_\beta(d,d';h_0)$ can be rewritten as

$$n[\hat{\Delta}(d,0) - \hat{\Delta}(d',0) + \triangle_*(d,d')]^T \Sigma_*(d,h)^{-1}[\hat{\Delta}(d,0) - \hat{\Delta}(d',0) + \triangle_*(d,d')][1+o_p(1)],$$

it follows from (4.27) that if $\triangle_*(d,d') = 0$, then $\max_{d,d' \in \mathcal{D}} |D_\beta(d,d';h_0)| = O_p(\log(N(\mathcal{D})))$ and

$$\exp(-D_\beta(d,d';h_0)/C_n) = \exp(-O_p(\log(N(\mathcal{D})))/C_n) = 1 + o_p(1).$$

However, if $\triangle_*(d,d') \neq 0$, then we have

$$D_\beta(d,d';h_0) = n||[\Sigma_*(d,h)]^{-1/2}[\triangle_*(d,d') + O_p(\sqrt{\log N(\mathcal{D})/n})]||_2^2.$$

Similarly, by following the proof of Theorem 1, we can prove similar results for $D_\beta(d,d';h)$, which finishes the proof of Theorem 2.

Proof of Corollary 1. For the sake of space, we only highlight the key difference between Theorem 1 and Corollary 1. It follows from Step (i) of Theorem 1 that for all $d \in \mathcal{D}$, we have

$$\hat{\beta}(d) - \beta_*(d) = M_n(\beta_*(d))^{-1} n^{-1} \sum_{i=1}^n g_{n,i}(\beta_*(d), d) + o_p(n^{-1/2}). \tag{4.31}$$

Let $P_1 = [I_{q_1} \vdots\ 0]$ and $P_2 = [0 \vdots\ I_{q_2}]$ be a $q \times q_1$ matrix and a $q_2 \times q$ matrix, respectively. Thus, we have $\hat{\beta}_2(d) - \beta_{*2}(d) = P_2 M_n(\beta_*(d))^{-1} n^{-1} \sum_{i=1}^n g_{n,i}(\beta_*(d), d) + o_p(n^{-1/2})$. Recall that $\hat{\beta}_1(d,h)$ is the solution to $\sum_{d' \in B(d,h)} \omega(d,d';h) \sum_{i=1}^n P_1 g_{n,i}((\beta_1(d), \hat{\beta}_2(d')), d') = 0$. We then use a Taylor's series expansion to show that

$$\begin{aligned}
&\hat{\beta}_1(d,h) - \beta_{*1}(d) = [\sum_{d' \in B(d,h)} \omega(d,d';h) P_1 M_n(\beta_*(d')) P_1^T]^{-1} \times \\
&n^{-1} \sum_{i=1}^n \sum_{d' \in B(d,h)} \omega(d,d';h) P_1 [I_q - M_n(\beta_*(d')) P_2^T P_2 M_n(\beta_*(d'))^{-1}] g_{n,i}(\beta_*(d'), d') \\
&+ o_p(n^{-1/2}).
\end{aligned}$$

Finally, we can approximate the covariance matrix of $\sqrt{n}\hat{\beta}_1(d,h)$ by using

$$[\sum_{d' \in B(d,h)} \omega(d,d';h) P_1 M_n(\beta_*(d')) P_1^T]^{-1} \times$$
$$n^{-1} \sum_{i=1}^{n} [\sum_{d' \in B(d,h)} \omega(d,d';h) P_1 [I_q - M_n(\beta_*(d')) P_2^T P_2 M_n(\beta_*(d'))^{-1}] g_{n,i}(\beta_*(d'),d')]^{\otimes 2} \times$$
$$[\sum_{d' \in B(d,h)} \omega(d,d';h) P_1 M_n(\beta_*(d')) P_1^T]^{-1}.$$

References

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

Abramowitz, M., and Stegun I., editor (1965). *Handbook of Mathmatical Functions.* New York: Dover Publications.

Alexander, D. C., and Barker, G. J. (2005). Optimal Imaging Parameter for Fibre-Orientation Estimation in Diffusion MRI. *NeuroImage*, **27**, 357-367.

Alexander, D. C., Barker, G. J., and Arridge, S. R. (2002). Detection and Modeling of Non-Gaussian Apparent Diffusion Coefficient Profiles in Human Brain Data. *Magnetic Resonance in Medicine*, **48,** 331-340.

Almli, C. R., Rivkin, M. J., and McKinstry, R. C. (2007). The NIH MRI study of normal brain development (Objective-2): newborns, infants, toddlers, and preschoolers. *Neuroimage*, **35,** 308-25.

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

Basser, P. J., Mattiello, J., and LeBihan, D. (1994 a). MR Diffusion Tensor Spectroscopy and Imaging. *Biophysical Journal*, **66,** 259-267.

Basser, P. J., Mattiello, J., and LeBihan, D. (1994 b). Estimation of the Effective Self-diffusion Tensor from the NMR Spin Echo. *Journal of Magnetic Resonance, Series B.*, **103,** 247-254.

Beckmann, C. F., Jenkinson, M. and Smith, S. M. (2003). General multilevel linear modeling for group analysis in fMRI. *NeuroImage*, **20**, 1052-1063.

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

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

Bowman, F. D. (2007). Spatio-temporal models for region of interest analyses of functional mappping experiments. *Journal of American Statistical Association*, **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.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman and Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B.*, **39,** 1-38.

Diggle, p., Heagerty, P., Liang, K. Y., and Zeger, S. (2002). Analysis of Longitudinal Data (2nd Ed.): Oxford University Press.

den Dekker, A. J., and Sijbers, J. (2005). Implications of the Rician Distribution for fMRI Generalized Likelihood Ratio Tests. *Magnetic Resonance Imaging* **23**, 953-959.

Escanciano, J. C. (2006). A Consistent Diagnostic Test for Regression Models using Projections. *Econometric Theory*, **22,** 1030-1051.

Evans, A. C., and B. D. C. Group. (2006). The NIH MRI study of Normal Brain Development. *NeuroImage*, **30,** 184-202.

Fan, Y., Batmanghelich, N., Clark, C. M., Davatzikos C. and the Alzheimers Disease Neuroimaging Initiative. (2008). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, **39**, 1731-1743

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

Fitzmaurice,G. M., Laird,N. M., and Ware, J. H. (2004). *SApplied Longitudinal Analysis.* John Wiley & Sons, New Jersey.

Gudbjartsson, H., and Patz, S. (1995). The Rician Distribution of Noisy MRI Data. *Magnetic Resonance in Medicine*, **34,** 910-914.

Haacke, E. M., Brown, R. W., Thompson, M. R., and Venkatesan, R. (1999). *Magnetic Resonance imaging: Physical Principles and Sequence Design.* New York: John Wiley and Sons.

Hardin, R. H., Sloane, N. J. A., and Smith, W. D. (1994). *Minimal Energy Arrangements of Points on a Sphere with Minimal 1/r Potential.* available at http://www.research.att.com/ njas/electrons/.

Henkelman, R. M. (1985). Measurement of Signal Intensities in the Presence of Noise in MR Images. *Medical Physics*, **12,** 232-233.

Hua, X., Lee, s., Yanovsky, I., Leow, A. D, Chou, Y., Ho, A. J.,Gutman, B., Toga, A. W., Jack, C. R., Bernstein,M. A., Reiman, E. M., Harvey, D. J, Kornak, J., Schuff, N., Alexander, G. E., Weiner, M. W., Thompson P. M. and the Alzheimer's Disease Neuroimaging Initiative. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: An ADNI study of 515 subjects. (2009). *NeuroImage*, **48.** 668-681

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

Jenkinson, M., Bannister, J. M., and Smith, S. M. (2002). Improved Optimisation for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, **17**, 825-841.

Jenkinson, M., and Smith, S. M. (2001). A Global Optimisation Method for Robust Affine Registration of Brain Images. *Medical Image Analysis*,**5**, 143-156.

Jones, D. K., and Basser, P. J. (2004). Squashing Peanuts and Smashing Pumpkins: How Noise Distorts Diffusion-Weighted MR Data. *Magnetic Resonance in Medicine*, **52**, 979-993.

Jones, D. K., Horsfield, M. A., and Simmons, A. (1999). Optimal Strategies for Measuring Diffusion in Anisotropic Systems by Magnetic Resonance Imaging. *Magnetic Resonance in Medicine*, **42,** 515-525.

Jorgensen, B. (1992). Exponential Dispersion Models and Extension: a Review. *International Statistical Review*, **60,** 5-20.

Karlsen, O. T., Verhagen, R., and Bovee, W.M. (1999). Parameter Estimation from Rician-distributed Data Sets using a Maximum Likelihood Estimator: Application to T_1 and Perfusion Measurements. *Magn. Reson. Med.*, **41,** 614-623.

Kochiyama, T., Morita, T., Okada, T., Yonekura, Y., Matsumura, M., Sadato, N. (2005). Removing the Effects of Task-related Motion using Independent-component Analysis. *NeuroImage*, **25**, 802-814.

Kristoffersen, A. (2007). Optimal Estimation of the Diffusion Coefficient from Non-averaged and Averaged Noisy Magnitude Data. *J. Magn. Reson.*, **187**, 293-305.

Lau, J. C., Lerch, J. P., Sled, J. G., Henkelman, R. M., Evans, A. C., Bedell, B. J. (2008). Longitudinal neuroanatomical changes determined by deformation-based morphometry in a mouse model of Alzheimer's disease. *NeuroImage*, **42,** 19-27.

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal Data Analysis using Generalized Linear models. *Biometrika*, **73,** 13-22.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221240.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society Series B* **44,** 190-200.

Luo, W. L., and Nicholas, T. N. (2003). Diagnosis and Exploration of Massively Univariate fMRI Models. *NeuroImage*, **19**, 1014-1032.

Macovski, A. (1996). Noise in MRI.*Magn. Reson. Med.* **36,** 494-497.

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

Meng, X. L., and Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, **80,** 267-278.

Nichols, T., and Hayasaka, S. (2003). Controlling the family-wise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, **12**, 419-446.

Nowark, R. D. (1999). Wavelet-based Rician Noise Removal for Magnetic Resonance Imaging. *IEEE Transactions on Image Processing* **8**, 1408-1419.

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. (2003). Image denoising: pointwise adaptive approach. *Annals of Statistics*, **31**, 30-57.

Polzehl, J. and Spokoiny, V. G. (2006). Propagation-separation approach for local likelihood estimation. *Probab. Theory Relat. Fields*, **135**, 335-362.

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

Qiu, P. (2007). Jump surface estimation, edge detection, and image restoration. *Journal of American Statistical Association*, **102**, 745-756.

Rice, S. O. (1945). Mathematical Analysis of Random Noise. *Bell System Technical Journal*, **24,** 46-156.

Rohde, G. K., Barnett, A. S., Basser, P. J., Marenco, S., and Pierpaoli, C. (2004). Comprehensive Approach for Correction of Motion and Distortion in Diffusion-weighted MRI. *Magnetic Resonance in Medicine*, **51,** 103-114.

Rowe, D. R. (2005). Parameter Estimation in the Magnitude-only and Complex-valued fMRI Data Models. *NeuroImage*, **25,** 1310-1324.

Rowe, D. R., and Logan, B. R. (2005). Complex fMRI Analysis with Unrestricted Phase is Equivalent to a Magnitude-only Model. *NeuroImage*, **24,** 603-606.

Salmond, C. H., Ashburner, J., Vargha-Khadem, F., Connelly, A., Gadian, D. G. and Friston, K. J. (2002). Distributional assumptions in voxel-based morphometry. *NeuroImage*, **17**, 1027-1030.

Scouten, A., Papademetris, X., Constable, R.T.(2006). Spatial resolution, signal-to-noise ratio, and smoothing in multi-subject functional MRI studies *NeuroImage*, **30,** 787-793

Sijbers, J., and den Dekker, A. J. (2004). Maximum Likelihood Estimation of Signal Amplitude and Noise Variance from MR Data. *Magnetic Resonance in Medicine*, **51,** 586-594.

Sijbers, J., den Dekker, A. J., Scheunders, P., and Van Dyck, D. (1998a). Maximum-likelihood Estimation of Rician Distribution Parameters. *IEEE Transactions on Image Processing*, **17,** 357-361.

Sijbers, J., den Dekker, A. J., Verhoye, M., Van Audekerke, J., and Van Dyck, D. (1998b). Estimation of Noise from Magnitude MR Images. *Magnetic Resonance Imaging*, **16,** 87-90.

Skare, S., Li, T., Nordell, B., and Ingvar, M. (2000). Noise Considerations in the Determination of Diffusion Tensor Anisotropy. *Magnetic Resonance Imaging*, **18,** 659-669.

Stute, W. (1997). Nonparametric Model Checks for Regression. *Annals of Statistics*, **25,** 613-641.

Styner, M., Lieberman, J. A., McClure, R. K., Weinberger, D. R., Jones, D. W., and Gerig, G. (2005). Morphometric analysis of lateral ventricles in schizophrenia and healthy controls regarding genetic and disease-specific factors. *Proceedings of the National Academy of Sciences USA*, **102**, 4872-4877.

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

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

Tuch, D. S., Reese, T. G., Wiegell, M. R., Makris, N., Belliveau, J. W., and Wedeen, V. J. (2002). High Angular Resolution Diffusion Imaging Reveals Intravoxel White Matter Fiber Heterogeneity. *Magnetic Resonance in Medicine*, **48,** 577-582.

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

van der Vaart (1998). *Asymptotic Statistics*. Cambridge Universirty Press.

van der Vaart and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

van der Vaart, A.W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* New York: Springer.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, **92**, 1-28.

Wei, B.C. (1998). *Exponential Family Nonlinear Models*. Singapore: Springer.

Weibull, A., Gustavsson, H., Mattsson, S., Svensson. J. (2008). Investigation of spatial resolution, partial volume effects and smoothing in functional MRI using artificial 3D time series. *NeuroImage*, **41,** 346-353

Whitwell, J. L. (2008) Longitudinal imaging: changes and causality. *Current Opinion in Neurology*. **21**, 410-416.

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

Xu, D., Mori, S., Solaiyappan, M., van Zijl, P. C., and Davatzikos, C. (2002). A Framework for Callosal Fiber Distribution Analysis. *NeuroImage*, **17,** 1131-1143.

Xue, Z., Shen, D., Karacali, B., Stern, J., Rottenberg, D. and Davatzikos, C. (2006). Simulating deformations of MR brain images for validation of atlas-based segmentation and registration algorithms. *NeuroImage*, **33**, 855-866.

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*, in press.

Zhu, H.T., Xu, D., Amir, R., Hao, X., Zhang, H., Alayar, K., Ravi, B., and Peterson, B. (2006). A Statistical Framework for the Classification of Tensor Morphologies in Diffusion Tensor Images. *Magnetic Resonance Imaging*, **24,** 569-582.

Zhu, H. T., Ibrahim, J. G., Tang, N., Rowe, D. B., Hao, X., Bansal, R. and Peterson, B. S. (2007a). A statistical analysis of brain morphology using wild bootstrapping. *IEEE Transaction on Medical Imaging*, **26**, 954-966.

Zhu, H.T., and Zhang, H.P. (2004). A Diagnostic Procedure based on Local Influence Measure. *Biometrika*, **91,** 579-589.

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). *Journal of the American Statistical Association*, **102**, 1085-1102.

Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.