Advanced Statistical Techniques in DW-MRI and fMRI Data Analysis

By

Allison Elisabeth Hainline

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

September 30, 2018

Nashville, Tennessee

Approved:

Jeffrey Blume, Ph.D.

Hakmook Kang, Ph.D.

Matthew Shotwell, Ph.D.

Bennett Landman, Ph.D.

This dissertation is dedicated to my late mother, Mary.
My loudest cheerleader, my guardian angel.
This one's for you, Mom!

ACKNOWLEDGEMENTS

Thank you to my advisor, Hakmook, for your support throughout my graduate career. You piqued my interest in medical imaging during my first visit to Vanderbilt and continued to help me grow in my knowledge through my dissertation. You are extremely thoughtful and always have my best interest at heart and I couldn't have asked for a better advisor.

A big thank you to my committee members in Biostatistics, Jeffrey and Matt. Each of you have contributed so much to my journey at Vanderbilt and I'm grateful for your influence. And to Bennett, a committee member and Research Assistantship supervisor: I could not have accomplished so much in the past two years without your guidance. The skills I learned as a member of the MASI lab are invaluable and continue to contribute to my success in new and exciting ways.

Thank you to the current and former members of the MASI lab at Vanderbilt. You took me under your wings when I knew *nothing* about image processing and stood with me as I learned the ins and outs of life in an engineering lab. The skills I have learned in MASI are unique and have fundamentally changed the way I conduct my research.

A special thank you to my family. To my parents, Charles and Mary, for always encouraging me to follow my dreams. The value you put on education was the spark that I trace all of my success back to. Thank you to my brother, Thomas, for setting such a high bar and instilling me with a love of learning. Your "did you know?"s taught me that learning can be fun (and also made me great at trivia!). And to my sister, Kelly, for bringing such joy to my life and teaching me to "bob". Our years together at Vanderbilt are so dear to me. I want to be like you when I grow up.

Finally, a resounding thank you goes out to my friends both within Biostatistics and without. You have all impacted me in varied and wonderful ways and I am so thankful for your friendship! Last, but not least, thank you to Kevin and Wallace for brightening every day and sticking with me through long hours of coding and writing (and complaining).

You're the best!

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

BOLD	Blood oxygen level-dependent
DLP	Dichotomous likelihood paradigm
DTI	Diffusion tensor imaging
DW-MRI	Diffusion-weighted magnetic resonance imaging
fMRI	functional magnetic resonance imaging
GFA	Generalized Fractional Anisotropy
HARDI	High Angular Resolution Diffusion Imaging
LP	Likelihood paradigm
MRI	Magnetic Resonance Imaging
ODF	Orientation Distribution Function
ROI	region of interest
SIMEX	Simulation Extrapolation
SNR	Signal-to-noise ratio

CHAPTER 1

INTRODUCTION

1.1 Brain Imaging

This dissertation aims to expand the understanding of the human brain via the application of modern statistical techniques for the analysis of existing image acquisition methods. Here we begin with a broad overview of neuroimaging as it relates to the work presented in this dissertation. This chapter concludes with an outline of the work proposed in Chapters 2-5.

1.1.1 Structural vs. functional imaging

Brain imaging can be separated into two broad areas: structural brain imaging and functional brain imaging. This dissertation addresses statistical problems in both. Modalities for structural brain imaging include computed axial tomography (CAT), magnetic resonance imaging (MRI) and positron emission tomography (PET), all of which allow for the discernment of brain structure that can be later used in disease diagnosis and other clinical applications. Modalities for functional brain imaging include PET, functional MRI (fMRI), electroencephalography (EEG) and magnetoencephalography (MEG), all of which can be used to study brain activity. In this dissertation, we will be focusing on two MRI modalities: diffusion-weighted MRI (DW-MRI) to study brain structure and fMRI to study brain function.

DW-MRI and fMRI differ in terms of the spatial and temporal resolution of each. All imaging modalities exhibit a unique balance between spatial resolution, temporal resolution, and degree of invasiveness. The spatial resolution of an image is what determines the clarity of the image, and thus, the ability of the experiment to distinguish changes between two spatial locations. The temporal resolution of an image is dependent on the speed at which images are acquired, and thus determines our ability to separate signal changes as a function of time. fMRI has high temporal resolution, which tends to come at the cost of lower spatial resolution due to the rapid succession of image acquisitions. DW-MRI acquisitions, however, are taken one image at a time, allowing for longer scan times, and resulting in higher spatial resolution. Both DW-MRI and fMRI are non-invasive, making them ideal modalities for both research and clinical applications.

Both structural and functional MRI images are acquired using the same type of scanner. In terms of the data structure, MRI images are often acquired in axial (horizontal) slices, which are stacked together to make up a 3D brain volume. This 3D volume is comprised of equally sized voxels (the 3D analog to a pixel), each of which corresponds to a unique spatial location within the volume. The size of these voxels is described by the spatial resolution of the scan. Voxels are the fundamental unit of both fMRI and DW-MRI.

1.1.2 MR physics

This section contains a very brief introduction to MR physics. This description is far from comprehensive, but aims to give the reader enough context to understand the proposed statistical techniques and their importance to the field.

The signal obtained in MRI is commonly derived from hydrogen atoms consisting of a single proton, which are found in water within the body tissues. These protons are positively charged and are always spinning, hence why they are often referred to as ‘spins’, which together generate a magnetic field. MRI measures the net magnetization of all hydrogen atoms within the volume of interest and exploits the magnetic properties of these protons in order to obtain signal from body tissues.

Each proton has an orientation in space, and, at rest, the orientations of each proton in a group will be completely random, thus cancelling one another out and leaving no net magnetization. The first step in MRI acquisition requires the application of a constant external magnetic field (B_0). This field causes the spins within the tissue to align in the direction of the applied field, resulting in a net longitudinal magnetization parallel to the field. At this point, each spin has a longitudinal component (M_z) that is aligned with the magnet, and a transverse component (M_{xy}) that rotates around the longitudinal component. The transverse components of the group of spins cancel each other out, while the longitudinal components exhibit a net magnetization in the opposite direction of the applied magnetic field.

In order to detect the signal, we need to be able to enhance the transverse signal, as any signal in the longitudinal direction will be drowned out by the strong magnetic field, B_0. To do this, we must apply an oscillating radio frequency (RF) pulse that serves to disrupt (or perturb) the aligned spins. The application of a 90° pulse will rotate the entire system of spins, resulting in a net transverse magnetization. At this point, the spins of the protons become in sync with one another, a state called phase coherence.

When the RF pulse is turned off, the system will return to its lowest energy (resting) state. In order for this to happen, energy must be transferred out of the system. The loss of transverse magnetization is referred to as transverse relaxation, and the regrowth of the longitudinal magnetization is referred to as longitudinal relaxation. During relaxation, the protons emit energy which makes up the signal, which is detected by a receiver coil.

The two types of relaxation are described by different time constants. Longitudinal relaxation is the restoration of magnetization along the longitudinal axis as the spins return to their resting state. This restoration is exhibited as exponential growth, and that exponential growth can be described by a time constant termed T1. T1 is the time it takes for the longitudinal magnetization to return to approximately 63% of the original net magnetization. Transverse relaxation is the loss of net transverse magnetization due to the loss of phase coherence. This loss is shown as exponential decay and is described by a time constant termed T2. T2 is the time it takes for the transverse magnetization to fall to 37% of its peak net magnetization.

Different tissue types have different T1 and T2 relaxation times. For example, CSF has a much longer T1 relaxation time than white matter. Image contrast allows for the control of which tissue characteristics are emphasized. This can be done by altering how often we excite the nuclei (relaxation time, or TR) and how soon after excitation we begin data collection (echo time, or TE). Long TR and short TE gives an idea of the proton density, long TR and long TE is a T2-weighted contrast, and short TR and short TE is a T1-weighted contrast. An additional contrast, T2*, is sensitive to flow and oxygenation and will be discussed further in Section 1.3.

1.2 Diffusion-weighted MRI

Diffusion-weighted MRI relies on the natural motion of water molecules within tissues. While not visible to the human eye, each molecule is in constant random motion as a result of thermal energy within the tissue, a phenomenon known as Brownian motion. Though the molecules are moving in a completely random manner, the diffusion of these molecules within tissues is impacted by the microstructure of the surrounding tissue. In the extracellular space, the water molecules are allowed to diffuse freely, however, the diffusion becomes restricted in intracellular space.

Thus, the water will diffuse along the path of least resistance, for example, around a cell rather than through the cell. This nature of water diffusion is the motivating factor for diffusion-weighted MRI, where we can exploit the natural motion of water molecules to infer the structure of the tissue on a microscopic scale. The result of a diffusion scan can also give indications of the structural changes of tissues of interest. For example, if the tissue is experiencing a great deal of cellular breakdown (necrosis), a scan may reveal an increased amount of diffusion in that area, as the diffusion becomes less hindered by cells.

Diffusion anisotropy describes the degree to which diffusion is hindered in one direction relative to another within a tissue. Unrestricted fluids which are allowed free diffusion in any direction are described as isotropic. Diffusion anisotropy appears when the microstructure is strongly aligned, as is the case in fibrous tissues. In this case, the diffusion of water within the tissue changes depending on the direction in which it is measured. In a fibrous tissue, such as a white matter, the diffusion is more restricted in the direction perpendicular to the tract than in the direction along the tract. This anisotropy provides a contrast mechanism for detecting the alignment of the material microstructure.

Diffusion-weighted image acquisition is most commonly performed using a pulsed-gradient spin echo (PGSE) sequence (Stejskal and Tanner (1965)). The pulsed field gradient allows for the measurement of the diffusion coefficient rather than the transverse relaxation. The first gradient dephases the magnetization and the second rephases it. If the molecules have diffused in the time between the pulses, they cannot be rephased and show up as a peak intensity decrease. The change in signal intensity depends on the rate of diffusion, the time of observation, and the strength of the gradient magnetic field.

1.2.1 Diffusion Tensor Imaging (DTI)

Diffusion tensor imaging (DTI) is a method for characterizing the microstructure of the brain for use in tracking changes due to disease or treatment. The diffusion tensor model was introduced by Basser et al. (1994*b*) and provided a systematic analytical framework for describing diffusion anisotropy in tissue. The diffusion tensor characterizes the magnitude, anisotropy and orientation of diffusion within the tissue. DTI estimates a diffusion tensor, and the eigenstructure of this tensor reveals the orientation of the diffusion compartments within the voxel. The major eigenvector is in the direction of the fiber orientation (Basser et al. (1994*b*)).

Once the diffusion tensor has been fit, several scalar measures can be calculated, each of which explains some property of the tensor. Of particular interest in this dissertation is the measure of fractional anisotropy (FA). FA is a dimensionless measure of the degree of anisotropy in diffusion (Basser and Pierpaoli (1996)). An FA of 0 represents unrestricted diffusion, while an FA of 1 implies that all of the diffusion occurs in a single direction, and that all other directions are restricted.

1.2.2 High Angular Resolution Diffusion Imaging (HARDI)

DTI assumes a 3-dimensional multi-variate Gaussian diffusion within each voxel (Alexander et al. (2007)). This assumption, however, may not apply in cases of restricted diffusion or partial volume effects, as is the case with multiple crossing or diverging fibers within a single voxel. To remedy this, Tuch et al. (2002) developed a diffusion imaging method that is able to measure the microscopic diffusion function within each voxel without requiring restrictive assumptions on the underlying diffusion function. High angular resolution diffusion imaging (HARDI) is able to discern crossing fibers via a high b-value diffusion gradient sampling scheme. High b-values are more effective in distinguishing between differing rates of diffusion of two fibers within the same voxel. HARDI obtains higher angular resolutions in order to reveal non-Gaussian diffusion. In HARDI, the diffusion gradient sampling is performed uniformly in 3 dimensions.

1.2.3 Q-ball

In this dissertation, we focus on a single model for reconstructing the HARDI signal: Q-ball imaging. Q-ball imaging was introduced by Tuch (2004) and utilizes the Funk-Radon transform to resolve multiple fibers within a single voxel without assuming a Gaussian diffusion process. Q-ball imaging allows for reconstruction of the diffusion orientation distribution function (ODF), which describes the probability density function of water diffusion along any direction.

In this dissertation, we use a regularized implementation of the Q-ball imaging approach proposed by Descoteaux et al. (2007) which uses a spherical harmonic basis along with a regularization term to simplify the Funk-Radon transform used in the original Q-ball fit. This regularized approach was shown to reduce ODF estimation errors and improve fiber detection while also providing a faster and more robust fit.

Without the diffusion tensor, we cannot calculate the fractional anisotropy within an MR acquisition. However, Tuch (2004) defines an extension of FA which uses the ODF to calculate a unitless, normalized measure of anisotropy, generalized fractional anisotropy (GFA).The methods proposed in Chapters 2-4 refer specifically to GFA as the metric of interest, though the theory should extend to other metrics calculated from the ODF.

1.3 Functional MRI

fMRI is a tool for uncovering brain function via an understanding of the dynamic changes in brain tissue that result from changes in neural metabolism and corresponding changes in blood flow dynamics (Chen and Glover (2016)). fMRI was introduced by three independent labs in 1992 (Bandettini et al. (1992); Ogawa et al. (1992); Kwong et al. (1992)). Nearly all fMRI experiments today rely on blood oxygen level-dependent (BOLD) changes in brain tissue.

BOLD imaging utilizes the different magnetic properties of oxygenated and deoxygenated hemoglobin in the blood in order to create a contrast that reveals changes in blood flow to brain regions. The blood in any region of the body has a local ratio of oxygenated hemoglobin (oxyhemoglobin) and deoxygenated hemoglobin (deoxyhemoglobin). Deoxyhemoglobin is paramagnetic, or weakly attracted to a magnetic field, while oxyhemoglobin is diamagnetic, or repelled by a magnetic field. These two different states of hemoglobin produce different local magnetic fields which impact the resulting signal in measurable ways.

In Section 1.1.2 we discussed T1 and T2 contrasts and how they relate different characteristics of tissue types. BOLD fMRI exploits an aspect of the T2 contrast called T2*. T2* is often referred to as the ‘observed’ T2. It turns out that in an MRI experiment, the transverse magnetization can decay faster than predicted as a result of local inhomogeneities in the surrounding tissue. Thus, the actual rate of transverse magnetization decay is referred to as T2*. In BOLD fMRI the local inhomogeneities that impact T2* are the magnetic properties of hemoglobin in its oxygenated or deoxygenated state. At rest, the blood flows at a normal rate and the T2*-weighted signal is normal. However, when the blood flow increases as a result of higher metabolic demands of the surrounding tissue, we see a decrease in the amount of deoxyhemoglobin, resulting in an increased T2*-weighted signal. This is due to the paramagnetic nature of deoxygenated hemoglobin, which distorts the magnetic field, causing a faster decay of the transverse magnetization, and thus a decrease in the time constant T2*. Deoxyhemoglobin suppresses the MR signal, thus as the concentration of deoxyhemoglobin decreases, the T2* signal increases. Alternatively, areas of the brain that have a higher concentration of oxyhemoglobin show a higher signal than those with more deoxyhemoglobin.

It is important to note that BOLD fMRI does not directly measure brain activation, but instead measures the metabolic demands (i.e. oxygen consumption) of active neurons. As the brain areas are activated, fresh (i.e. oxygenated) blood is sent to the area, resulting in a higher concentration of oxyhemoglobin, and thus, a higher signal. While the modality is not perfect, simultaneous fMRI and electrophysiological studies have confirmed that the BOLD contrast mechanism does reflect neural response to a stimulus. However, much more research is yet to be done to truly understand how much of the neuronal response can truly be understood via BOLD contrast (Logothetis and Wandell (2004)).

1.3.1 Resting-state fMRI

The methodology proposed in Chapter 5 of this dissertation pertains exclusively to resting-state fMRI. Correlated fluctuations in a resting brain were described by Biswal et al. (1995). In this work, the authors concluded that the correlation of these fMRI signals was evidence of brain activity even in the absence of an experimental task. Since then, resting-state fMRI (rs-fMRI) has been used to identify resting state networks (RSNs), which define networks of brain regions that demonstrate synchronous activation without a task or stimulus, the most popular of which is the default mode network (DMN) (Lee et al. (2013)).

By definition, all resting-state data sets are acquired at a baseline state, i.e. subjects are instructed to do nothing but lay still in the scanner. With the exception of the lack of specific tasks, resting-state fMRI data are acquired in the same way as task-induced fMRI, as described in Section 1.3. rs-fMRI is generally focused on identifying low-frequency BOLD fluctuations (between 0.01 and 0.08 Hz), in contrast to task-induced fMRI which tends to focus on higher frequency signals.

Clinically, rs-fMRI has been used to distinguish between the resting brain patterns in a variety of psychiatric and neurological disorders, such as Alzheimer's (Li et al. (2002); Wang et al. (2006)), multiple sclerosis (Lowe et al. (2008, 2002)), schizophrenia (Zhou et al. (2008)).

1.3.2 Analysis

The analysis of fMRI data poses many statistical challenges due to the very large nature of the data. Because each brain volume has approximately 100,000 voxels, all of which have been measured across a range of time periods, for multiple subjects, any voxel-based analysis becomes a challenging statistical problem. Considering the weak nature of the signal of interest along with a complicated temporal and spatial noise structure, makes it clear that this type of data analysis requires a specific set of techniques tailored to these characteristics.

1.3.2.1 Preprocessing

Before we can begin analyzing the data, several preprocessing steps must be performed. The exact preprocessing techniques can vary widely from study to study, but most include a selection of the same common steps: slice-time correction, motion correction, co-registration, normalization, spatial smoothing and low-pass filtering. These steps aim to decrease or eliminate any contamination in the data that is not related to spontaneous neural activity, i.e. subject motion, respiratory and cardiac effects, and hardware artifacts (Weissenbacher et al. (2009)). Specific approaches for each preprocessing step are reviewed in detail by Chen and Glover (2016).

The pre-processing steps of rs-fMRI data follows the same general steps as in task-related fMRI with the addition of a low-pass filter that retains frequencies less than 0.08 Hz. This filtering helps separate out cardiac and respiratory effects that occur at higher frequencies, and to improve the signal-to-noise ratio (Lee et al. (2013)).

1.3.2.2 Functional connectivity

Functional connectivity (FC) is defined as the undirected association between two or more fMRI time series. FC relates brain regions to one another functionally, but makes no assumptions about the structural connection between the brain regions.

In this dissertation, we focus on bivariate connectivity, which provides information about relationships between pairs of regions. To learn about bivariate connectivity, we first calculate the cross-correlation between the time series taken from two brain regions of interest. This correlation is often transformed into z-scores using Fisher's z-transformation. The correlations can be calculated for each subject individually, followed by a group analysis of all subjects. In this dissertation, the group analysis is performed by first calculating the pairwise correlations, applying Fisher's z-transformation and conducting a t-test for each pair of brain regions.

1.3.2.3 Multiple comparisons

Due to the number of pairwise comparisons that are made for an ROI-based rs-fMRI analysis, the concept of multiple comparisons correction becomes extremely important. Multiple comparisons corrections come in two major categories: family-wise error rate corrections and false discovery rate corrections.

The family-wise error rate (FWER) is the probability of making at least one type I error across the entire "family" of hypothesis tests performed. Methods that control the family-wise error rate include Bonferroni, random field theory, and permutation tests. These methods provide strong control over the number of false positives, but can tend to be conservative, leading to decrease in power.

The Bonferroni correction aims to control the FWER by directly manipulating the significance threshold. If P_i is the p-value from a test of hypothesis H_i, then we reject H_i if $P_i \leq \frac{\alpha}{m}$, where m is the total number of hypotheses, and α is the significance level. This method assumes that individual p-values are independent, which is violated in fMRI data with spatial correlation.

Another popular method for FWER correction is a permutation testing technique. Permutation tests are nonparametric tests that rely on the assumption that, under the null hypothesis, the data labels are exchangeable. To perform a permutation test, first create permutations of the original observed data and calculate a t-statistic for each permuted data set. Together, these t-statistics make up an empirical cumulative distribution function (CDF) of the test statistic under the null. The p-value is then calculated by finding the probability of observing a test statistic that is at least as extreme as the test statistic calculated from the observed data (Nichols and Holmes (2001)).

The low power of FWER correction methods when errors are correlated leads many researchers to utilize a different technique for multiple testing: false discovery rate correction. The false discovery rate (FDR) is the expected value of the proportion of false positives among the total number of positive test results. Controlling the FDR ensures that, *on average*, the FDR is no larger than some pre-specified rate, q. The Benjamini-Hochberg procedure is one of the most popular for FDR control and involves specifying the FDR q and ranking the p-values from the family of tests from smallest to largest. For m comparisons, we define k as the largest i such that $P_{(i)} \leq \frac{i}{m}q$ and reject all $P_{(1)}, P_{(2)}, \ldots, P_{(k)}$ (Benjamini and Hochberg (1995)). This method is particularly useful since it can be applied to any valid statistical test, as it works on the p-values, rather than the test statistics themselves.

The analysis described in Chapter 5 demonstrates both permutation testing and FDR correction as baseline techniques to compare performance with the proposed method.

1.4 Dissertation Focus

As neuroimaging studies become more numerous and data are increasingly available, the need for improved understanding of the statistical properties of such data increases as well. There is room for specific techniques for understanding these statistical properties as well as for utilizing them to allow for proper statistical inference.

The dimensionality of both DW-MRI and rs-fMRI data complicates the analysis and forces researchers to develop new techniques to account for it. In particular, the rs-fMRI community needs improved approaches for dealing with multiple comparisons in such a way that maintains high power while not compromising on Type I error control.

1.4.1 Open problems

The field of medical imaging is progressing very rapidly, but gaps remain particularly in the area of statistical analysis of such images. Image processing involves the calculation of several voxel-wise metrics from data acquisitions. However, these metrics are often considered absolutes, rather than as statistics with their own distributional properties. This is likely because well known ways of calculating these distributions are not accessible. Further, publicly available data sets are widely available, encouraging the combination of seemingly similar studies in an effort to increase statistical power. However, very little research has been done in understanding the impact of such data combination. There exists no agreed-upon way for assessing each data set in terms of its statistical and noise properties and thus ensuring increased power as a result of increased sample size. Finally, frequentist methods in fMRI data analysis tend to fall in one of two extremes: either the analyses end up being far too conservative or they fail to correct for multiple comparisons and have the opposite problem.

1.4.2 Contributions

- We propose statistical approaches for estimating bias and variance from a single HARDI acquisition that are useful for bias-correction, data quality assurance, and data combination and meta-analysis (Chapter 2).

- We provide an application of the bias and variance estimation methods from Chapter 2 to an evaluation of inter-site bias and variance in traveling subjects. This includes a workflow with advice for utilizing our methods in evaluating data quality (Chapter 3).

- In order to speed up the bias and variance estimation process, we develop a deep learning extension of the aforementioned statistical approach that allows for the estimation to be performed up to 200× faster (Chapter 4).

- We introduce a novel inferential method for functional magnetic resonance imaging (fMRI) studies via an application of the Likelihood paradigm that allows for simultaneous control of type I and type II error rates and shows superior performance to popular frequentist analysis techniques (Chapter 5).

CHAPTER 2

EMPIRICAL SINGLE SAMPLE QUANTIFICATION OF BIAS AND VARIANCE IN Q-BALL IMAGING

2.1 Introduction

Diffusion-weighted magnetic resonance imaging (DW-MRI) is an image acquisition technique that utilizes the natural diffusion of water within the human body to non-invasively study tissue microstructure. High angular resolution diffusion imaging (HARDI) is able to identify crossing fiber orientations, while diffusion tensor imaging (DTI) is unable to discern fibers in more than a single direction.

The accuracy, precision, and sensitivity of DTI have been studied and noise has been shown to have an impact on the bias and variance of DTI-derived metrics, such as FA (Lauzon et al. (2011, 2013); Basser (1997); Bastin et al. (1998); Skare et al. (2000); Basser and Pajevic (2000); Anderson (2001); Chang et al. (2007); Farrell et al. (2007); Hutchinson et al. (2017)). Previous work has applied the statistical concept of simulation extrapolation (SIMEX) (Carroll et al. (1996); Cook and Stefanski (1994)) in an effort to quantify the bias in an observed data acquisition (Lauzon et al. (2011, 2013)). This work extends the application of SIMEX to a single, empirical HARDI acquisition fit with a Q-ball model.

Bias and variance play a critical role in any statistical analysis. The appropriate balance between bias and variance allows for optimal information gain. It is not enough to have an unbiased estimator (i.e., an estimator that estimates the correct quantity, on average) if the variance is comparatively large (i.e., the estimate is imprecise). Similarly, an estimate with very small variance but large bias is very precise, but will not converge to the correct value as the sample size increases. Each situation can be useful, depending on the goals of the study, but a balance between the two is considered optimal. Thus, it is important to evaluate the methods that we use in terms of the bias and variance of the estimates we produce. The ability to quantify the bias and variance between different imaging and analysis methods quickly and easily allows the researcher to make informed design decisions at every step. Currently, there are no methods available for bias correction in HARDI acquisitions. This chapter provides approaches for the quantification of both bias and variance that can be performed with a single data acquisition, allowing for the comparison across methods without requiring repeated data acquisitions.

In addition to the simple quantification of bias for better understanding of the quality of a variety of imaging acquisition parameters and model fitting strategies, this method may also prove useful as a tool for bias correction. Several acquisition parameters are set during the image acquisition process, all of which can affect scan quality and the resulting analysis. Of particular interest is the situation where two scans are taken at different acquisition settings on the same subject. In theory, scans should be comparable within subjects, but in practice, they can be extremely variable. The SIMEX process aims to provide a method for bias-correction while the bootstrap procedure allows for an empirical estimate of the variance of Q-ball metrics, such as generalized fractional anisotropy (GFA), which is a measure of anisotropy calculated from the orientation distribution function (ODF).

This work appeared in Hainline, Nath, Parvathaneni, Blaber, Schilling, Anderson, Kang and Landman (2018).

2.2 Theory

This analysis focuses on Q-ball imaging reconstruction of the ODF of HARDI data acquisitions. We evaluate methods detailed by (Hess et al. (2006)) and (Descoteaux et al. (2007)) that use a spherical harmonic basis in the reconstruction of the ODF, which describes the patterns of diffusion within the tissue.

The regularized Q-ball reconstruction introduced in (Descoteaux et al. (2007)) requires the $m \times 1$ vector of diffusion weighted signals at each voxel, as well as a $2 \times m$ matrix of gradient directions in spherical coordinate space, where m is the number of gradient directions. Details on the calculation of the ODF are found in (Descoteaux et al. (2007)).

Following calculation of the ODF, a variety of metrics can be calculated. The metric used in this analysis is GFA, which is given by,

$$GFA = \frac{std(\psi)}{rms(\psi)} = \sqrt{\frac{m \sum_{i=1}^{m} (\psi_i - \overline{\psi})^2}{(m-1) \sum_{i=1}^{m} \psi_i^2}}$$

where ψ is the ODF vector, and $\overline{\psi}$ is its mean (Tuch (2004)).

2.2.1 SIMEX applied to empirical data

The SIMEX approach detailed herein estimates the bias present in metrics generated by a Q-ball model. SIMEX was introduced as a method for performing inference in models where measurement error was a concern. This method is unique in that it does not require fitting a complicated parametric measurement error model, yet still provides unbiased estimates of the object of interest. In short, bias estimates can be calculated by adding increasing amounts of synthetic noise to the data and computing the desired metrics at each noise level. The resulting measures can be extrapolated to the hypothetical case where no noise is present (Cook and Stefanski (1994)).

In the case where the measurement error variance can be well estimated, SIMEX is able to estimate the bias without requiring any model fitting, as is generally the case in typical measurement error estimation and correction procedures. SIMEX requires that the measurement error variance is well estimated and that the metric of interest changes monotonically as a function of noise added.

In this section, we adopt the notation established by Lauzon et al. (2011). We assume X_{truth} to be a truth dataset with zero noise. This dataset can be used to calculate the ground truth metric, GFA_{truth}, via a Q-ball fit. Now, assume an observed dataset with experimental noise where σ_E represents the standard deviation of the noise at each voxel. This observed dataset is given by,

$$\mathbf{X}_{obs} = \mathbf{X}_{truth} + \eta_{\sigma_E}$$

which represents the addition of stacked Rician noise with standard deviation σ_E as in Lauzon et al. (2013). While the original SIMEX approach assumes a normally distributed noise term, we are able to substitute the Rician distribution due to its approximation of a Gaussian distribution at SNR greater than 3 (Gudbjartsson and Patz (1995)). The metrics resulting from a Q-ball fit of this observed data are given by GFA_{obs}.

The “simulation” portion of the SIMEX procedure begins with the simulation of noisy GFA observations. These noisy observations cannot be simulated directly, thus a series of Monte Carlo simulations are performed via the addition of stacked Rician noise with the standard deviation of $\omega^{\frac{1}{2}}\sigma_E$, given by,

$$\mathbf{X}_{M.C.}(\omega) = \mathbf{X}_{obs} + \eta_{\omega^{\frac{1}{2}}\sigma_E}$$

where M.C. represents a Monte Carlo replication and metrics derived from $\mathbf{X}_{M.C.}(\omega)$ are given by $GFA_{M.C.}(\omega)$. The average value within each set of replications within a given ω value is given by $\overline{GFA}_{M.C.}(\omega)$.

Once a trend is established in the sequence of $\overline{GFA}_{M.C.}(\omega)$ values, a quadratic equation is fit.

$$GFA = \beta_0 + \beta_1\omega + \beta_2\omega^2;\ \omega = 0, 1, 2, \ldots$$

The variance of the simulated, noisy data can be viewed as a function of ω,

$$var(\mathbf{X}_{M.C.}(\omega)) = var\left(\eta_{\sigma_E} + \omega^{\frac{1}{2}}\eta_{\sigma_E}\right) = \sigma_E^2(1+\omega)$$

Thus, the variance goes to zero when $\omega = -1$ and the value of GFA at this value is considered noiseless. The value at $\omega = -1$ is referred to as GFA_{SIMEX}.

Finally, the bias can be estimated by subtracting the SIMEX metrics from those calculated from the observed data:

$$\widehat{Bias} = GFA_{obs} - GFA_{SIMEX} \tag{2.1}$$

The true bias can be calculated, given truth data, by

$$Bias_{true} = GFA_{obs} - GFA_{truth} \tag{2.2}$$

2.2.2 Bootstrap estimation of variance

The bootstrap is a popular method for estimation of variance in models where a parametric solution is either not available or not assumed. The family of bootstrap procedures involves the resampling of data with replacement, where each resampled dataset is viewed as a surrogate for an independently sampled dataset. The metric of interest is computed for each resampled dataset. Through repeated sampling, we can create a bootstrap distribution of the metric, which approximates the true sampling distribution of the metric (Efron (1992)).

In a case where multiple acquisitions were obtained for each gradient direction, a classic bootstrap approach works well. In a classic bootstrap, data points are resampled across the acquisitions, but within each voxel. The requirement for repeated acquisitions, however, is unreasonable for larger numbers of gradient diffusion directions. The residual bootstrap, in which residuals are resampled at random across data points, cannot be used with a single DW-MRI acquisition due to the heteroscedasticity (i.e. non-constant variance across gradient directions) that is introduced upon permutation (Basser et al. (1994*a*); Whitcher et al. (2008)). While the signal used for Q-ball does not use a log-transform (as in DTI), we may still observe some heteroscedasticity as a result of differences in the variance of the Rician distribution as a function of the mean.

In cases where the errors are heteroscedastic, the wild bootstrap has proven effective at characterizing the variance without assuming constant variance (Liu (1988)).

This method was applied to DTI in (Whitcher et al. (2008)) and was shown to outperform the regular bootstrap for a variety of settings. The wild bootstrap has since been applied in probabilistic fiber tracking with positive results (Jones (2008)). For these reasons, we chose to employ the wild bootstrap for our variance estimation procedure as follows.

A wild bootstrap technique gives empirical estimates of the variance of popular metrics from a Q-ball fit. Continuing the notation established above, we begin with the truth data, X_{truth} with zero experimental noise. n datasets are simulated via the addition of Rician noise with standard deviation σ_E. These datasets are seen as hypothetical independent acquisitions from a single subject. These datasets are fit using the Q-ball and GFA for each voxel is calculated. The standard deviation of the resulting calculations gives σ_{true}.

Next, variance is estimated via the wild bootstrap procedure. First, residuals must be calculated at each voxel,

$$\epsilon = \mathbf{X}_{truth} - \mathbf{X}_{obs} \tag{2.3}$$

In order to create the bootstrapped datasets, the signs of the residuals are flipped randomly and added back to the observed data (Jones (2008)). The signs of the residuals are determined by a vector of random Bernoulli draws of the same length as the residual vector:

$$\mathbf{X}_{boot} = \mathbf{X}_{obs} + \epsilon B \tag{2.4}$$

where $\mathbf{X}_{obs}$ is the $m \times 1$ vector of observed data at a single voxel and B is an $m \times 1$ vector of random Bernoulli draws. This process is repeated n times, resulting in n simulated datasets. GFA is calculated for each simulated dataset and the standard deviation of GFA across the n datasets gives $\hat{\sigma}$.

2.3 Methods

A flowchart detailing the steps of the SIMEX process on GFA as well as the calculation of bootstrap variance is given in Figure 2.1. The SIMEX procedure is performed on a voxel-by-voxel basis, where each voxel is evaluated independently. All calculations were performed in Matlab version R2016a (MATLAB (2016)) and the Camino Diffusion MRI toolkit (Cook et al. (2006)).

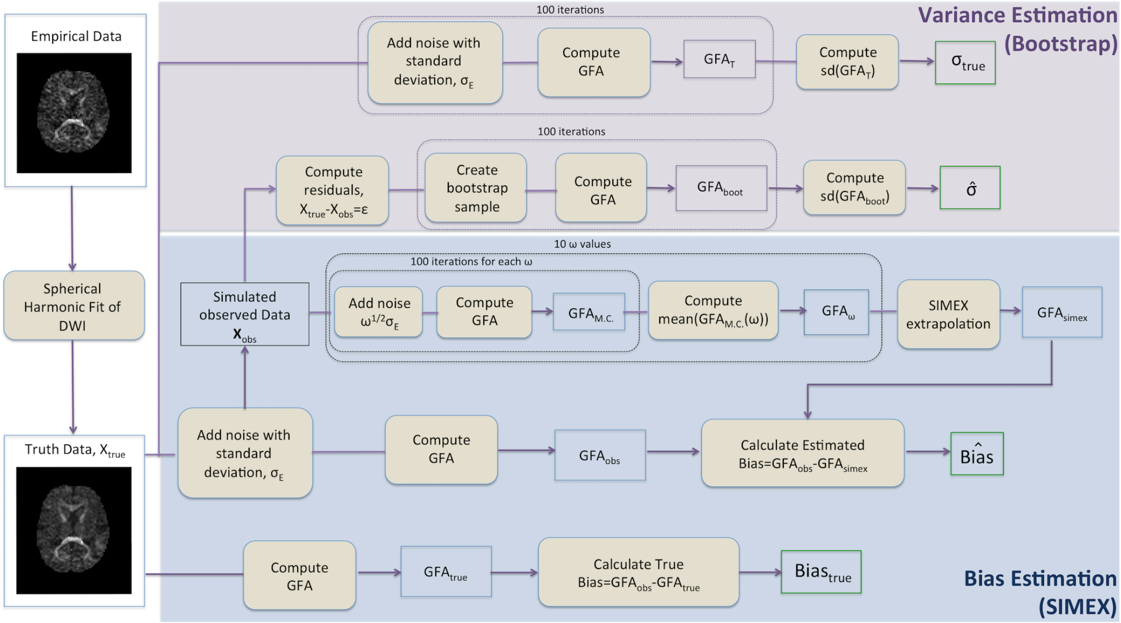
Empirical Data
Spherical Harmonic Fit of DWI
Truth Data, X_{true}
Variance Estimation (Bootstrap)
100 iterations
Add noise with standard deviation, σ_E
Compute GFA
GFA_T
Compute sd(GFA_T)
σ_{true}
100 iterations
Compute residuals, X_{true}-X_{obs}=ε
Create bootstrap sample
Compute GFA
GFA_{boot}
Compute sd(GFA_{boot})
$\hat{\sigma}$
10 ω values
100 iterations for each ω
Simulated observed Data $\mathbf{X}_{obs}$
Add noise $\omega^{1/2}\sigma_E$
Compute GFA
$GFA_{M.C.}$
Compute mean($GFA_{M.C.}$(ω))
GFA_ω
SIMEX extrapolation
GFA_{simex}
Add noise with standard deviation, σ_E
Compute GFA
GFA_{obs}
Calculate Estimated Bias=GFA_{obs}-GFA_{simex}
$\hat{Bias}$
Compute GFA
GFA_{true}
Calculate True Bias=GFA_{obs}-GFA_{true}
$Bias_{true}$
Bias Estimation (SIMEX)

Figure 2.1: Flowchart detailing the SIMEX and bootstrap procedures. The flowchart begins with the creation of the true data via a spherical harmonic fit of the empirical data as detailed in the text. The top two sections refer to the calculation of the estimated and true standard deviation for GFA. The calculation of true standard deviation is based on the true data while the calculation of the estimated standard deviation is based on the noise-added simulated observed data. The SIMEX procedure begins with the simulated observed data, iterates through a series of noise values and concludes with the extrapolation to estimate GFA. True bias requires the calculation of GFA based on the truth data and comparison with the observed GFA as calculated from the simulated observed data.

2.3.1 Empirical data

The empirical data used in this experiment were obtained from a healthy volunteer using a 3T Philips Scanner with a 32-channel head coil. The session consisted of 96 gradient directions at a b-value of 3000 s/mm^2. The voxel resolution is 2.5mm x 2.5mm x 2.5mm with 38 slices. The scan parameters were: Multi-Band=2; SENSE=2.2; TR= 2650 ms; TE=94 ms; partial Fourier=0.7. Fold over direction was A-P with a P (posterior) fat shift. For each shell, an additional diffusion scan was acquired with reverse phase encoded volumes (i.e., fold over direction A-P with A fat shift) with a minimally weighted volume and 3 diffusion weighting directions with a b-value of 1000 s/mm^2 along the imaging frame cardinal directions, and all other parameters were kept constant. All data were acquired in accordance with the Vanderbilt University Institutional Review Board (IRB) guidelines and with the signed consent of the volunteer.

2.3.2 Creating ground truth data

To create the ground truth data used in this experiment, the 6^{th} degree spherical harmonic basis function was selected. This basis function was generated using the b-vectors used in the original data acquisition. Regularized linear least squares fitting was used to estimate the spherical harmonic coefficients of the diffusion-weighted signal for each voxel. The resulting spherical harmonic series representation is a smoothed version of the original data; thus the resulting brain volume is assumed to be noiseless and is used as the ground truth dataset for the entirety of this analysis. Note that this method depends heavily on the appropriateness of the Q-ball model. Any deviation from the model may result in a systematic bias that cannot be corrected via the SIMEX procedure. Care must be taken when fitting the truth model to avoid such bias. This dataset, along with the b-values and b-vectors can be found here: www.nitrc.org/projects/masimatlab under "SIMEX on HARDI."

2.3.3 Creating observed data

The observed data, $\mathbf{X}_{obs}$, were created via the addition of random Rician noise to the ground truth data. The standard deviation of the Rician noise is the standard deviation of the residuals, σ_E, thus this observed dataset approximates an empirically observed HARDI data acquisition at the given SNR.

2.3.4 SIMEX

2.3.4.1 Calculating estimated bias

The first step of the SIMEX process is to calculate $\overline{GFA}_{M.C.}(\omega)$. 100 Monte Carlo simulations were performed for each $\omega = 1, 2, \ldots, 10$, and the average was taken for each ω to obtain $GFA_{M.C.}(\omega)$. A quadratic equation was fit in order to extrapolate to the GFA value that results when $\omega = -1$. The resulting value is known as GFA_{SIMEX}. GFA_{obs} was obtained directly from a Q-ball fit of the observed data, $\mathbf{X}_{obs}$. Estimated bias was calculated as the difference between GFA_{SIMEX} and GFA_{obs} (Eq. 2.1).

2.3.4.2 Calculating true bias

In order to calculate the true bias, the GFA of $\mathbf{X}_{obs}$ was calculated via a Q-ball fit. GFA_{truth} was calculated by fitting X_{truth} with no additional noise. True bias was calculated by taking the difference between GFA_{obs} and GFA_{truth} (Eq. 2.2). An example of the SIMEX procedure on 3 voxels from different brain regions is shown in

Figure 2.2. A successful SIMEX procedure is one where the SIMEX estimated GFA value is closer to the true GFA value than the observed GFA value, i.e., the estimated bias is close to the true bias.

This procedure was performed independently for each voxel in the brain volume.

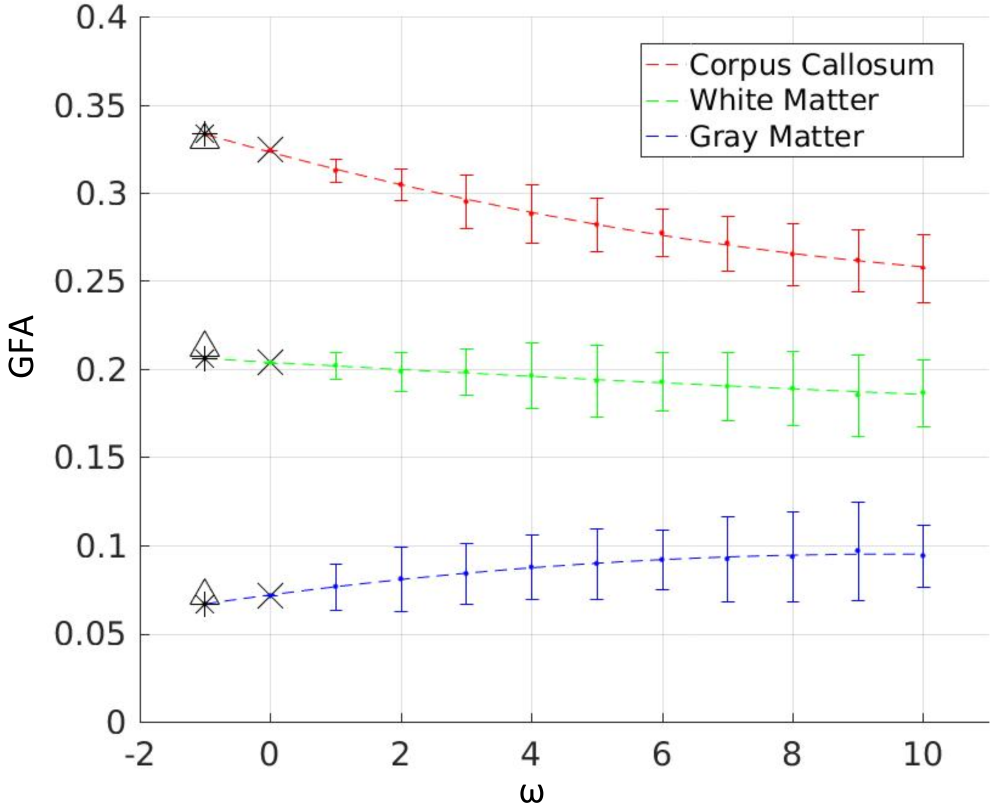

Figure 2.2: The SIMEX procedure demonstrated on 3 distinct voxels within the brain. These three voxels show the three possible results of the SIMEX procedure: when the SIMEX estimate improves the observed estimate, when the SIMEX estimate is approximately equal to the observed estimate, and when the SIMEX estimate is worse than the observed estimate. ω is the multiplier on the amount of noise added to each voxel (note that $\omega = 0$ is the observed data and $\omega = -1$ is the noiseless true data). Each $\bullet$ indicates the mean of the 100 Monte Carlo iterations at that ω value. The $\times$ indicates the observed GFA value for that voxel. The triangles indicate the true GFA value for each voxel. The asterisks indicate the SIMEX estimated GFA value for each voxel. The error bars represent the 5^{th} and 95^{th} percentiles of the M.C. iterations for each ω.

2.3.5 Bootstrap

2.3.5.1 Obtaining residuals

To obtain the necessary residuals for the residual wild bootstrap, we calculate the difference in signal between the ground truth data and the observed data (Eq. 2.3).

2.3.5.2 Calculating estimated variance

We obtain an estimate of the bootstrap variance for GFA. Residual wild bootstrap is performed using X_{boot} and ϵ (Eq. 2.4). We obtain 100 GFA estimates by repeating the bootstrap procedure and calculating GFA for each residual bootstrap sample. The standard deviation of the 100 metrics is taken to obtain the estimated standard deviation of the procedure, $\hat{\sigma}$.

2.3.5.3 Calculating true variance

Starting with the true data, X_{truth}, we add random Rician noise with standard deviation, σ_E, to obtain an observed dataset, X_{obs}. The Q-ball model is fit and GFA_{obs} is obtained. This process is repeated 100 times, resulting in 100 GFA values. The standard deviation of the 100 GFA values is considered to be the true standard deviation of the procedure, σ_{true}.

2.3.6 Characterization on independent datasets

We have applied the methods described above on two additional datasets to evaluate the generalizability of the approach. The second dataset is from the 2017 ISMRM TraCED challenge (https://my.vanderbilt.edu/ismrmtraced2017/). These data were acquired on a 3T Philips scanner and consisted of 64 gradient directions at a b-value of 3000 s/mm^2 with a voxel resolution of 2.5mm x 2.5mm x 2.5mm with 44 slices. The scan parameters were: Multi-Band=2; SENSE=2.2; TE=99 ms; partial Fourier=0.755. Fold over direction was A-P with a P (posterior) fat shift.

The third dataset we used is from the 2015 ISMRM Tractography challenge (Maier-Hein et al. (2017)). These data were obtained from an artificial phantom that was generated using the Fiberfox software. The anatomy was based on bundles segmented from a Human Connectome Project subject. These data are a 2mm isotropic diffusion acquisition with 32 gradient directions at a b-value of 1000 s/mm^2. An artifact-free ground truth dataset was also provided and was used as the truth dataset in this analysis.

These two additional datasets were analyzed in the same fashion as the first, as detailed in the Methods section, with one exception for the 2015 ISMRM Tractography challenge data. Since the 2015 ISMRM Tractography challenge included a truth dataset, we opted to use it rather than the Q-ball fitted model as X_{truth}.

2.4 Results

2.4.1 Simulation results

Figure 2.3 displays the results of the SIMEX and bootstrap procedures and their ability to estimate the true bias and standard deviation of the empirical data. Performance was evaluated on white matter and gray matter separately. The SIMEX procedure tends to underestimate the true bias of GFA in cases where the true bias is larger. The bootstrap procedure is successful at estimating the true standard deviation of GFA in both white matter and gray matter.

Figure 2.4 provides a qualitative look at the performance of SIMEX and the wild bootstrap on GFA. In this figure, we compare the true GFA values with their estimated values and compute the difference between the two, or the residual bias of our methods. Also included for reference are maps of the observed GFA and a B0 image as a structural reference. In both cases, the procedures appear to estimate their targets well. These qualitative results are in line with the quantitative results shown in Figure 2.3.

2.4.2 Sensitivity to noise

We have also provided the error of the SIMEX procedure as a function of SNR in Figure 5a. The root mean squared error (RMSE) decreases as the SNR increases for both white matter and gray matter. We have identified a typical clinical SNR range between 20:1 and 40:1 and find that the SIMEX method works well within this range.

The SIMEX procedure (Figure 2.5 (panel a)) shows a 5-7% improvement over the uncorrected estimates of GFA in white matter and a 5-8% improvement over the uncorrected estimates of GFA in gray matter, within the SNR range. At lower SNR, the bias-correction procedure shows minimal improvement due to the noisy nature of the data. The largest improvements are seen in the meaningful range, though the procedure continues to result in lower RMSE at SNR up to 70:1.

To evaluate the performance of the bootstrap variance estimation procedure as a function of SNR, we examined the ratio of the estimated standard deviation and the true standard deviation for white matter and gray matter (Figure 2.5 (panel b)). We find that within the SNR range, the wild bootstrap procedure is able to capture 97% of the true standard deviation for white matter and 86% for gray matter.

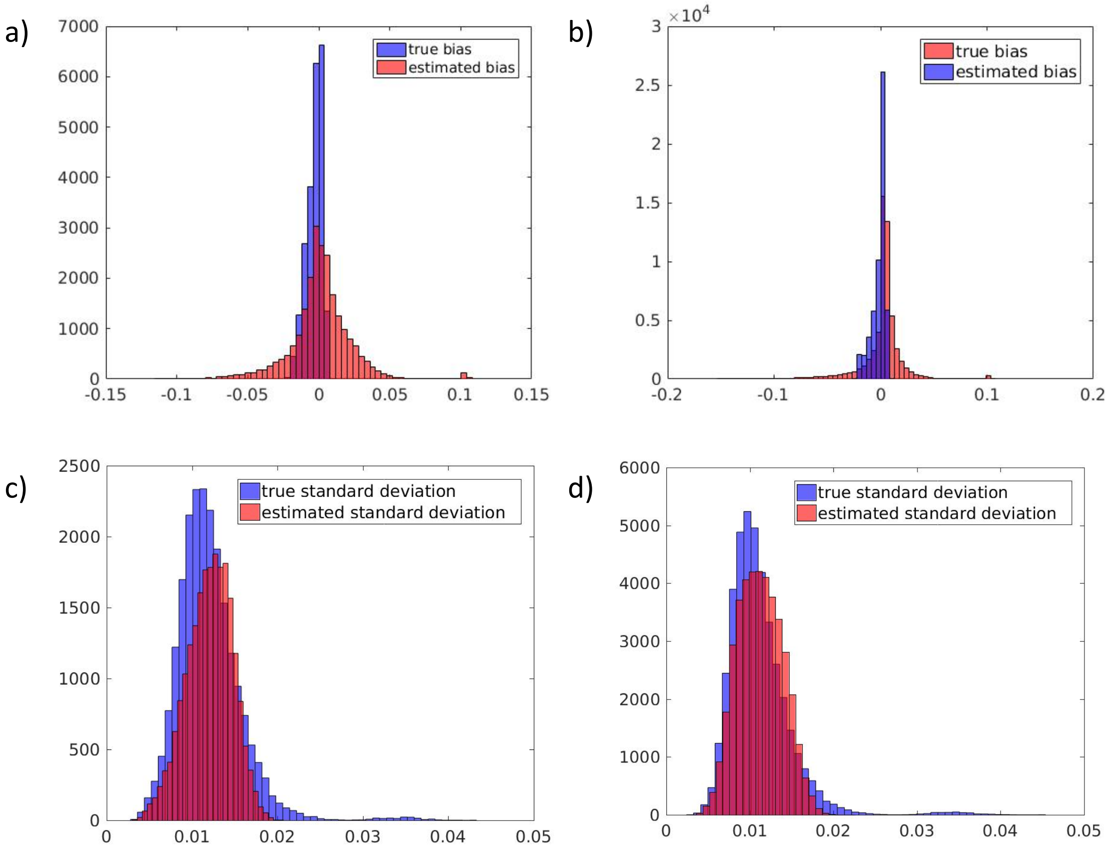

Figure 2.3: Bootstrap approximations of the variance of GFA. a) True bias (blue) and estimated bias (red) in white matter. b) True bias and estimated bias in gray matter. c) True standard deviation and estimated standard deviation in white matter. d) True standard deviation and estimated standard deviation in gray matter. The SIMEX procedure appears to overestimate the true bias in white matter and underestimate the true bias in gray matter. The bootstrap procedure well characterizes the variance.

2.4.3 Performance on independent datasets

For the 2017 ISMRM TraCED challenge dataset, the SIMEX procedure (Figure 6a) shows 5-6% improvement over the uncorrected estimates of GFA in white matter and a 5-7% improvement over the uncorrected estimates of GFA in gray matter, within the SNR range. We also find that within the SNR range, the wild bootstrap procedure is able to capture about 95% of the true standard deviation for both white matter and gray matter (Figure 6b).

For the 2015 ISMRM Tractography challenge dataset, the SIMEX procedure (Figure 6c) shows 3-11% improvement over the uncorrected estimates of GFA in white matter and 5-8% improvement over the uncorrected estimates of GFA in gray matter, within the SNR range. The wild bootstrap procedure performs poorly on the 2015 ISMRM Tractography challenge dataset, overestimating the true standard deviation

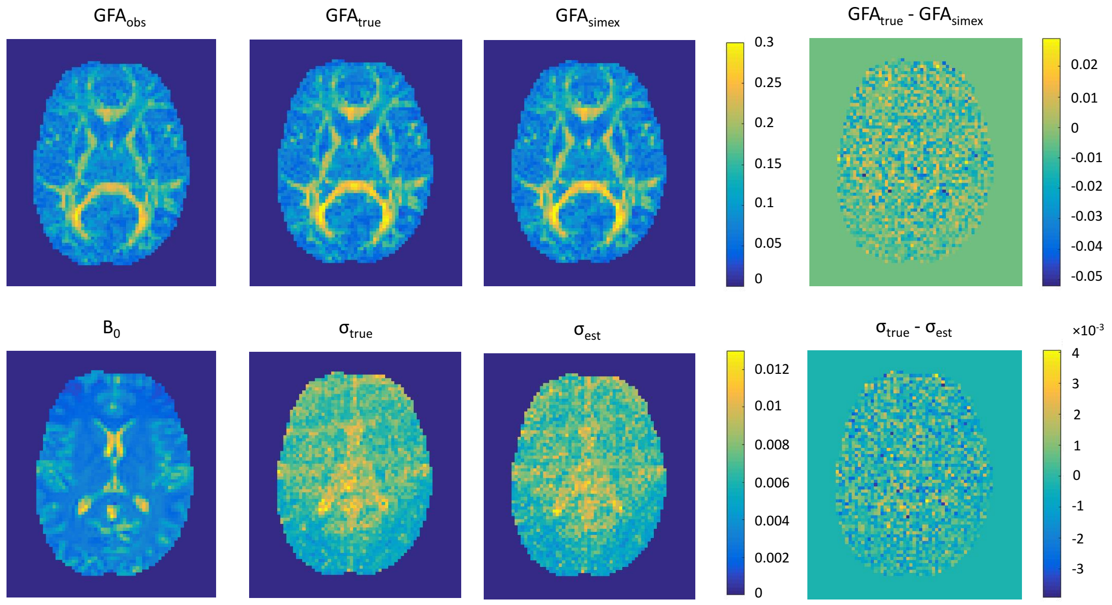

Figure 2.4: Qualitative results demonstrate the performance of SIMEX bias-correction as well as the wild bootstrap standard deviation estimation on GFA measures at an SNR of 20:1. The approaches described here can closely estimate the true bias and standard deviation from single acquisition data with no repeated volumes. In both cases, the magnitude of the difference between the estimated and true values is several times smaller than that of the original measures (note the scales of the color bars for the difference images), thus we are able to accurately estimate both the bias and variance with these techniques. The B_0 image is provided as a structural reference for the variance estimation results.

by 20-40% in white matter and 10-12% in gray matter (Figure 6d). The use of the provided truth data as the X_{truth} rather than the Q-ball model as we did for the other two datasets negatively impacts the performance of the wild bootstrap.

2.5 Discussion

The interpretation of DW-MRI imaging is highly dependent on the conditions and parameters involved in the image acquisition. In addition, systematic bias has the potential to mislead the results of a statistical analysis of imaging data. Biased measurements in a diagnostic setting can have a negative impact on treatment decisions, while bias in a research setting may mislead methodological comparisons and imply false hierarchies among analysis methods. The ability to correct the bias of an empirical sample without requiring a parametric model fit is valuable. With the methods described here, we can evaluate each acquisition independently, allowing for the identification of imaging artifacts and other data quality issues on a case-by-case

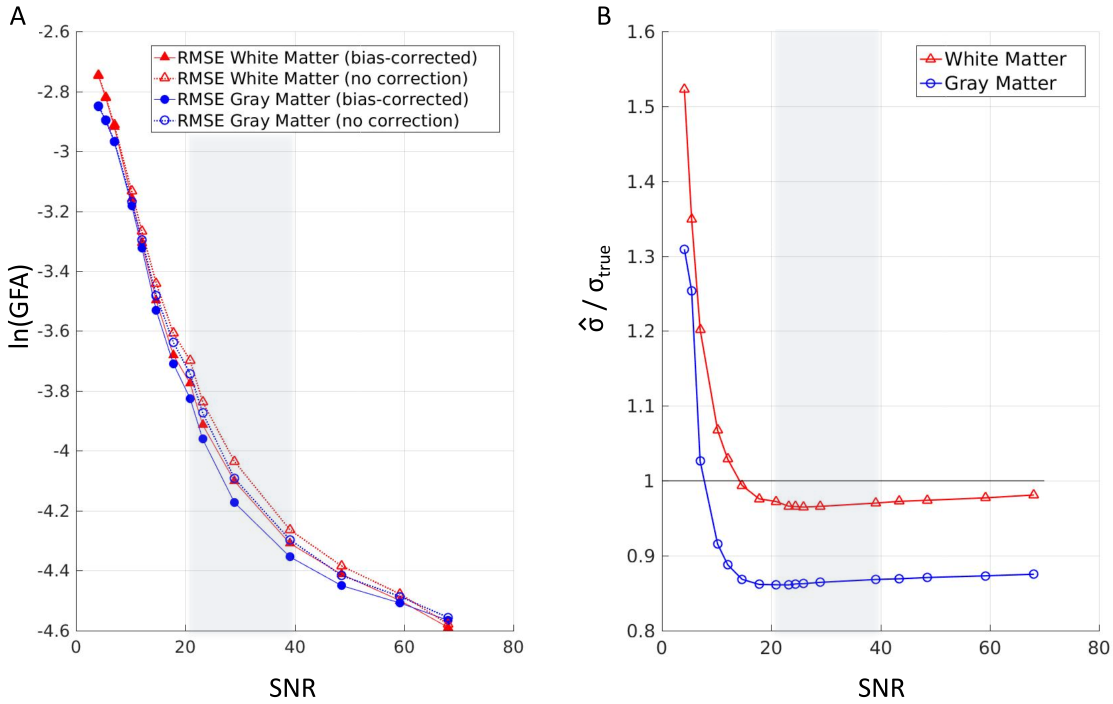

Figure 2.5: Performance of SIMEX and bootstrap across a range of SNR. (a) The root mean squared error (RMSE) of the SIMEX GFA estimation (log-scale) after bias-correction shows improvement over the RMSE without bias-correction. Within the meaningful SNR range (shown in the gray box), the bias-corrected estimates show a 5-7% improvement over the uncorrected estimates in white matter and a 5-8% improvement over the uncorrected estimates in gray matter. Within the range, lower SNR shows greater improvements. (b) The ratios of mean estimated standard deviation of GFA and mean true standard deviation of GFA within white matter and gray matter are shown across a range of SNR values. We find that the wild bootstrap procedure slightly underestimates the standard deviation in white matter, where the estimates are about 97% of the true values in our specified meaningful SNR range. In gray matter, the underestimation is larger, though the bootstrap procedure still captures 86% of the true variation. Within the typical clinical SNR range, the methods estimate the standard deviation well. Below this range, we see lower performance and above this range, we see small improvements.

basis.

Bias correction is an important factor for single scans as well as for repeated scans as part of a longitudinal study of a single patient, however, there is no accepted method available for bias estimation in HARDI data. The application of the SIMEX bias-correction technique can prove useful in such studies, where changes within subject are of interest. In addition, the ability to quantify the bias and variance of a single scan proves useful for the comparison of different scan settings or analysis methods. Comparison of bias and variance among analysis methods allows for a more comprehensive summary of the usefulness of each and may better inform the choice of method for future studies.

The methods described in this chapter are conceptually simple and computationally feasible, making them excellent candidates for inclusion in standard data processing procedures. The speed of these procedures is dependent on the time taken to add random noise to the data as well as the time taken to fit the model and compute the metric of interest. The SIMEX methodology is flexible in terms of the number of ω values as well as the extrapolation function. These design considerations should be made based on the behavior of the metric of interest. The number of ω values is chosen depending on the metric and should be large enough to capture the trend of the noise-added metrics, but not so large that variation from one value to the next is lost. The user may also choose the degree of the polynomial fit to the noise-added metrics, though we have found that a quadratic fit tends to work well for most cases and reduces the possibility of overfitting.

The SIMEX method detailed herein assumes that the noise has a Rician distribution with mean zero. While data acquisitions are unlikely to have a truly zero mean noise distribution, the bias is small in high SNR data leading us to maintain this assumption for simplicity. A possible limitation of this technique lies in the assumption that the Q-ball model is correct and does not introduce any systematic bias to the procedure. The reliance on a model is not unique to this procedure and care should always be taken in fitting an appropriate model in order to limit the amount of bias introduced to the analysis. Additionally, potential issues for the application to human data include increased noise levels or imaging artifacts. The SIMEX and bootstrap approaches are not guaranteed to work well in the presence of extreme imaging artifacts and such artifacts may have negative impacts on the resulting analysis. Care should be taken in the preprocessing steps to eliminate any sources of imaging noise that may affect the analysis. Due to the procedure's dependence on noise for bias estimation, we have found that the performance requires a relatively high SNR in order to perform optimally. Extensions to this method that may improve the performance at lower SNR will be the focus of our future work.

Perhaps the most interesting results are those seen in Figure 6. In validating our work with additional datasets, we found that the use of the spherical harmonic Q-ball model as $\mathbf{X}_{truth}$ is crucial to the ability of the wild bootstrap to accurately estimate the true standard deviation of the GFA estimates. The wild bootstrap relies on the residuals between $\mathbf{X}_{truth}$ and $\mathbf{X}_{obs}$ and when the two come from different models, this model mismatch dominates the Monte Carlo procedure's performance, leading to inaccurate estimates of the standard deviation of GFA. It is encouraging to note that the SIMEX procedure is unaffected by the choice of truth model and thus proves itself to be a general-purpose technique for bias estimation in HARDI data acquisitions.

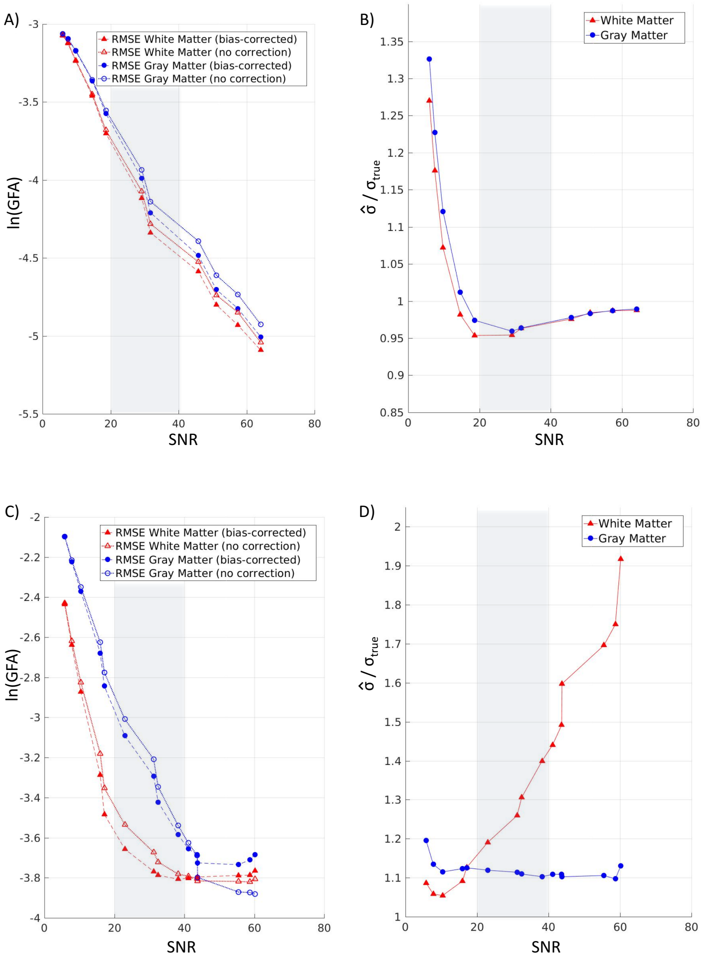

Figure 2.6: Performance of SIMEX and bootstrap across a range of SNR for two additional datasets, 2017 TraCED challenge and 2015 Tractography challenge. (a) For the TraCED challenge, the root mean squared error (RMSE) of the SIMEX GFA estimation (log-scale) after bias-correction shows improvement over the RMSE without bias-correction. Within the meaningful SNR range (shown in the gray box), the bias-corrected estimates show a 5-6% improvement over the uncorrected estimates in white matter and a 5-7% improvement over the uncorrected estimates in gray matter. Within the range, lower SNR shows greater improvements. (b) For the TraCED challenge, the ratios of mean estimated standard deviation of GFA and mean true standard deviation of GFA are shown across a range of SNR values. We find that the wild bootstrap procedure slightly underestimates the standard deviation in both white matter and gray matter, where the estimates are about 95% of the true values in our specified meaningful SNR range. Within the typical clinical SNR range, the methods estimate the standard deviation well. (c) For the 2015 Tractography challenge, the root mean squared error (RMSE) of the SIMEX GFA estimation (log-scale) after bias-correction shows improvement over the RMSE without bias-correction. Within the meaningful SNR range (shown in the gray box), the bias-corrected estimates show a 3-11% improvement over the uncorrected estimates in white matter and a 5-8% improvement over the uncorrected estimates in gray matter. (d) For the 2015 Tractography challenge, the ratios of mean estimated standard deviation of GFA and mean true standard deviation of GFA are shown across a range of SNR values. We find that the wild bootstrap procedure overestimates the standard deviation in both white matter and gray matter, where the estimates are 20-40% higher than the true values in our specified meaningful SNR range for white matter, and 10-11% higher in gray matter. We find that when there is a model mismatch, as is the case with our use of the Tractography dataset, the wild bootstrap technique cannot accurately estimate the true standard deviation.

CHAPTER 3

EVALUATION OF INTER-SITE BIAS AND VARIANCE IN DIFFUSION-WEIGHTED MRI

3.1 Introduction

Given pragmatic considerations of study design and magnetic resonance imaging (MRI) data acquisition, many clinical studies combine data from several different sources in order to increase the sample size and improve the power (Di Martino et al. (2014); Jack et al. (2014)). Traditional techniques for evaluating contrasts and testing differences across time assume that the bias and variance are constant across all acquisitions (e.g., Basser and Jones (2002)). However, there is no universal technique for evaluating sources of bias and variance in MRI on individual subjects. Violation of statistical assumptions has the potential to invalidate inferences. Thus, the addition of new data has the potential to decrease the statistical power as a result of the introduction of bias. Significant amounts of bias and variance can result even within a single site due to patient factors, hardware differences, and signal processing/software.

Here, we focus on the context of high angular resolution diffusion imaging (HARDI), with a specific focus on Q-ball imaging (QBI). To illustrate the problem, Figure 3.1 presents the variation that can be observed within a single subject scanned across 5 separate scans (3 independent scanners and 2 re-scans). Each of these scans was taken under comparable acquisition parameters and should reveal the same brain structure; however, the figure shows variation in B0, DWI, and vector-mapped images. These differences are clearly visually appreciated from the images themselves, but the extent of the variation's impact on HARDI analysis is difficult to quantify visually. In diffusion tensor imaging (DTI), bias and variance haven been assessed for single subjects with simulation extrapolation (SIMEX) and Monte Carlo methods, respectively (Lauzon et al. (2013)). Recently, these methods have been adapted to HARDI (Hainline, Nath, Parvathaneni, Blaber, Schilling, Anderson, Kang and Landman (2018)), but have not been evaluated on multi-site traveling data. This chapter presents an analysis of 3 sites using harmonized HARDI acquisition protocols. The focus of this work is to consider tools for estimating the bias and variance, as well as to present a decision tree to guide the researcher as to how such data could be used. The process described herein can be used for quality assurance within a single site to ensure optimal statistical power and results.

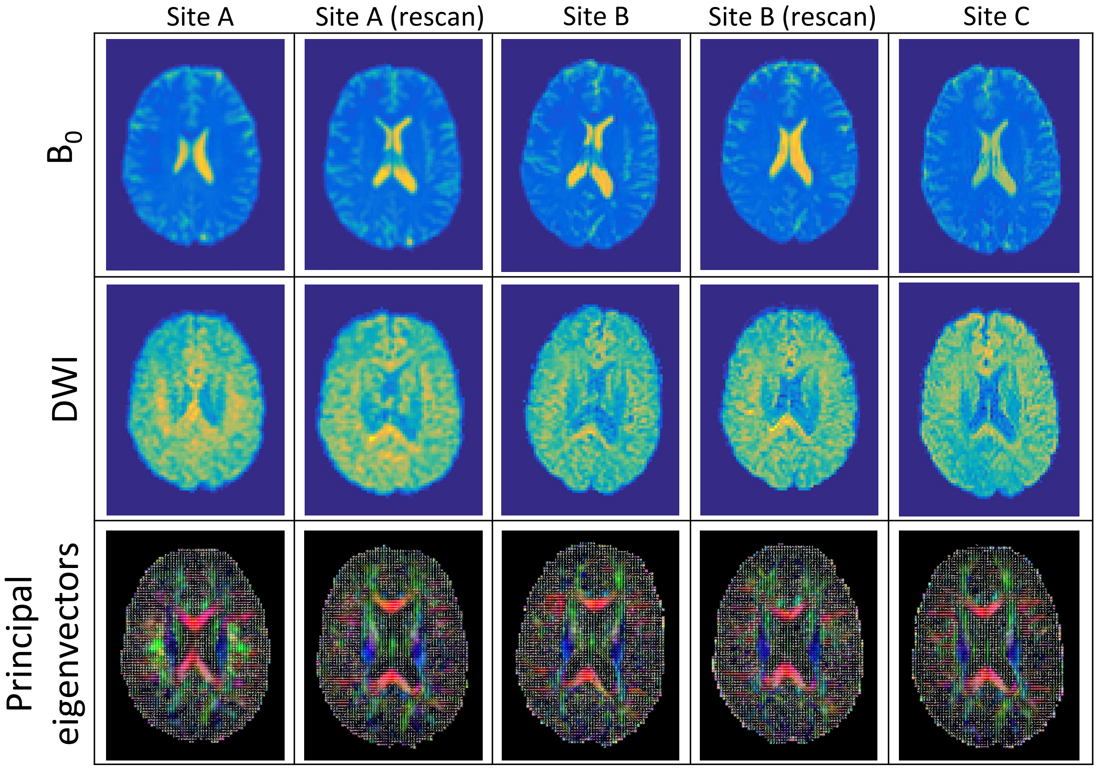

Figure 3.1: Illustration of the variation that may be seen across scan sites within a single subject with each scan processed in the acquired space. B0, DWI, and the principal eigenvectors for the same mid-axial slice across different scans within the same subject. The color intensity maps are constant across all sites for each (arbitrary units).

This work appeared in Hainline, Nath, Parvathaneni, Blaber, Rogers, Newton, Luci, Edmonson, Kang and Landman (2018).

3.2 Methods

3.2.1 Data acquisition

Subjects were imaged at 3 independent study sites. Five subjects were imaged at Site A on a 3.0T system using a full body transmit coil with a 32-channel head only receive coil. Non-diffusion weighted imaging sequences consisted of 3D T1 weighted MPRAGE, resting state fMRI, and B0 mapping. Diffusion weighted imaging sequences consisted of a 96 direction DTI (b=1000, 1500, 2000, 2500 s/mm^2; SENSE = 2.5; partial Fourier factor = 0.77; voxel dim. = 1.9x1.9mm^2; FOV = 112x112; # of slices = 48, slice thickness = 2.5mm), as well as regularly interspersed acquisitions of a three direction DWI acquisition acquired with reversed phase encoding gradients, and finally a vendor standard 30 direction DTI (b=1000 s/mm^2) acquisition.

Scans from this site were resampled to 2.5x2.5mm^2 for comparison with the other two study sites. 4 subjects were re-scanned with the same protocol. Five subjects were imaged at Site B on a 3.0T system using a full body transmit coil with a 32-channel head only receive coil. Non-diffusion weighted imaging sequences consisted of 3D T1 weighted MPRAGE, resting state fMRI, and B0 mapping. Diffusion weighted imaging sequences consisted of a 96 direction DTI (b=1000, 1500, 2000, 2500 s/mm^2; SENSE = 2.5; partial Fourier factor = 0.77; voxel dim. = 2.5x2.5mm^2; FOV = 96x96; # of slices = 48, slice thickness = 2.5mm), as well as regularly interspersed acquisitions of a three direction DWI acquisition acquired with reversed phase encoding gradients, and finally a vendor standard 30 direction DTI (b=1000 s/mm^2) acquisition. 4 subjects were re-scanned with the same protocol. Four subjects were imaged at Site C on a 3.0T system using a full body transmit coil with a 32-channel head only receive coil. Non-diffusion weighted imaging sequences consisted of 3D T1 weighted MPRAGE, resting state fMRI, and B0 mapping. Diffusion weighted imaging sequences consisted of a 96 direction DTI (b=1000, 1500, 2000, 2465 s/mm^2; GRAPPA = 2; voxel dim. = 2.5x2.5mm^2; FOV = 96x96; # of slices = 50, slice thickness = 2.5mm), as well as regularly interspersed acquisitions of a three direction DWI acquisition acquired with reversed phase encoding gradients, and finally a vendor standard 30 direction DTI (b=1000 s/mm^2) acquisition.

3.2.2 Model fitting

We fit all data with a Q-ball imaging with a model order 6 reconstruction of the orientation distribution function (ODF) of the HARDI data acquisitions. As detailed in Hess et al. (2006) and Descoteaux et al. (2007), we use a regularized spherical harmonic reconstruction of the ODF and calculate generalized fractional anisotropy (GFA),

$$GFA = \frac{std(\psi)}{rms(\psi)} = \sqrt{\frac{m\sum_{i=1}^{m}(\psi_i - \overline{\psi})^2}{(m-1)\sum_{i=1}^{m}\psi_i^2}}$$

where ψ is the ODF vector, and $\overline{\psi}$ is its mean (Tuch (2004)).

Note that we have chosen to use GFA as the metric of interest for this analysis. However, the theory behind these methods allows for calculations based on any scalar metric, with the only requirements being that it is continuous and monotonic with respect to the addition of noise (Cook and Stefanski (1994)).

3.2.3 GFA bias estimation

We used the SIMEX approach to estimate the bias of GFA. This approach was adapted from modern statistical methods (Cook and Stefanski (1994)) and has been described in Lauzon et al. (2013) for use in DTI, and for HARDI in Hainline, Nath, Parvathaneni, Blaber, Schilling, Anderson, Kang and Landman (2018). When measurement error is present, the true data are unable to be observed, i.e., we instead observe

$$\mathbf{X}_{obs} = \mathbf{X}_{truth} + \eta_{\sigma_E}$$

which represents the addition of stacked Rician noise (R) with standard deviation σ_E as in Lauzon et al. (2013).

In short, the SIMEX procedure relies on the behavior of the metric of interest as a function of the addition of random noise. As noise is added to the observed data in increasing amounts, a trend is observed in the metric of interest calculated at each level of noise. We can use this trend to extrapolate backward to the case with no measurement error (or noise) and obtain a function of Cook and Stefanski (1994).

SIMEX does not require the fitting of parametric measurement error models in order to estimate the bias (Carroll et al. (1996)). The only requirements for the application of SIMEX are that the measurement error variance can be estimated and that the metric of interest is smooth and monotonic as a function of noise.

3.2.4 GFA variance estimation

The wild bootstrap was used to estimate the variance of GFA. The wild bootstrap, as detailed in Jones (2008), is a method for estimating the variance of an MRI-derived metric, without requiring the use of several repeated data acquisitions. We use the wild bootstrap rather than a traditional bootstrap resampling with replacement due to the heteroscedasticity of the errors in a diffusion model (Basser et al. (1994*a*); Whitcher et al. (2008)).

3.2.5 Analysis

For this analysis, we have data acquired from two shells, b=1000 and b=2500 s/mm^2. These shells will be analyzed separately and results are compared. The SIMEX and bootstrap procedures are performed on a voxel-by-voxel basis, where each voxel is evaluated independently. All calculations were performed in Matlab version R2016a (MATLAB (2016)) and the Camino Diffusion MRI toolkit (Cook et al. (2006)).

The first step in the analysis is the estimation of both the bias and variance (as detailed in sections 2.3 and 2.4) for each voxel within each data acquisition. For evaluation, all data was registered to the ICBM 2009a Nonlinear Symmetric template (Fonov et al. (2009, 2011)). We chose to do our analysis on the ROI level; thus, we used manually delineated five white matter ROIs: centrum semiovale, splenium of the corpus callosum, internal capsule, putamen, and globus pallidus. The average bias and variance values were taken within ROIs, resulting in a value for each subject, scanner, b-shell, and ROI combination.

When the average values for the bias and standard deviation of GFA are calculated for each ROI, the decision tree found in Figure 3.2 can be used to guide the model selection process. First, we create quantile-quantile (Q-Q) plots and histograms for both the average bias and average standard deviation of GFA values within each ROI. Q-Q plots compare the quantiles of the observed distribution to that of a Gaussian distribution. Ideally, the points on the Q-Q plots fall directly on the line, though slight deviations are often not cause for alarm. Often, deviations that occur in the tails of the distribution may be due to small sample sizes and are expected to stabilize with larger samples. Note that formal tests of normality are available; however, these tests can be misleading (Mason and Schuenemeyer (1983)). Thus, we recommend visually checking for normality via the Q-Q plots and histograms. Bimodality or other signs of asymmetry will be evident in both types of plots. If the plots reveal non-normality, the recommendation is to closely examine the raw data for artifacts and correct them, if found. If no artifacts are found, non-parametric methods that do not assume normality should be used.

Next, boxplots are made for the mean values of the bias and standard deviation of GFA for each ROI with the points overlaid. Each point represents the average value within the ROI for either the estimated bias of GFA or the estimated standard deviation of GFA. The boxplot allows for the visual identification of outliers and gives a clear picture of the differences across sites and subjects. This boxplot is the main tool for determining what modeling strategy is optimal for the data distributions.

3.3 Results

This analysis consisted of 5 subjects who were scanned at up to 3 independent scan sites with re-scans. Our first step was to create bias and standard deviation maps, shown in Figure 3.3 for subject 01. These maps help reveal the spatial distributions

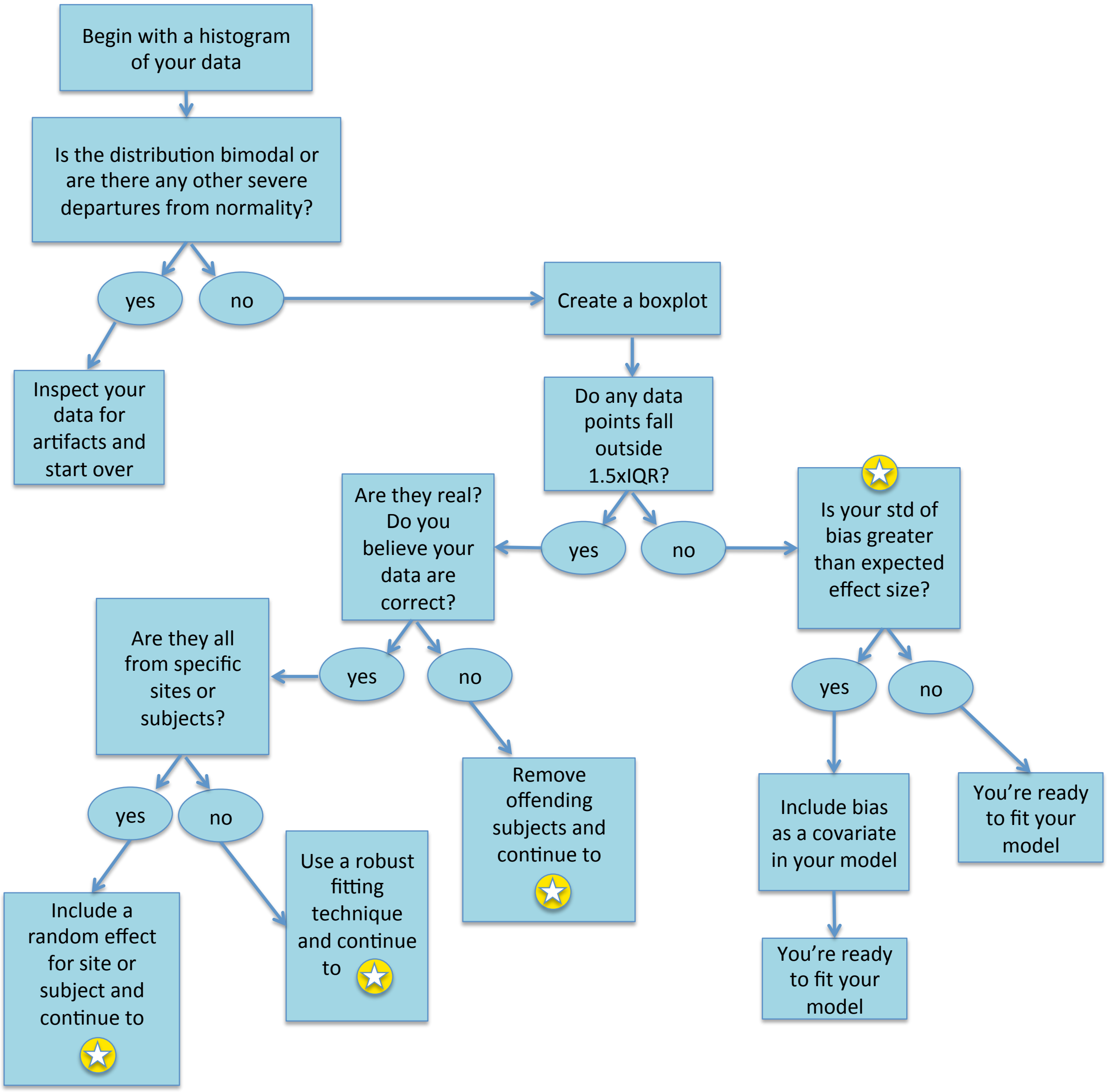
Begin with a histogram of your data
Is the distribution bimodal or are there any other severe departures from normality?
yes
no
Inspect your data for artifacts and start over
Create a boxplot
Do any data points fall outside 1.5xIQR?
yes
no
Are they real? Do you believe your data are correct?
Is your std of bias greater than expected effect size?
yes
no
Are they all from specific sites or subjects?
Remove offending subjects and continue to
yes
no
Include bias as a covariate in your model
You're ready to fit your model
yes
no
Include a random effect for site or subject and continue to
Use a robust fitting technique and continue to
You're ready to fit your model

Figure 3.2: Recommended decision-making process for model selection. The procedure begins with a simple histogram or density plot to identify any severe distributional issues that would interfere with inference. For the purpose of this decision tree, the "data" refers to the averaged values within each ROI for either the estimated bias of the metric of interest or the estimated standard deviation of the metric of interest. Once the distribution has been checked, a boxplot is recommended to get an idea of what outliers exist and where they came from. Any patterns in outliers that can be attributed to site- or subject-specific artifacts should be accounted for with a random effect. Finally, one should look at the standard deviation of the metric as well as the magnitude of the bias. If either of these is larger than the expected effect size, this should be accounted for in the modeling through the inclusion of a covariate

of both the bias and the variance across the brain structures. In particular, we see higher bias and standard deviations in the gray matter in comparison to white matter.

Following the decision tree in Figure 3.2, we made Q-Q plots and histograms

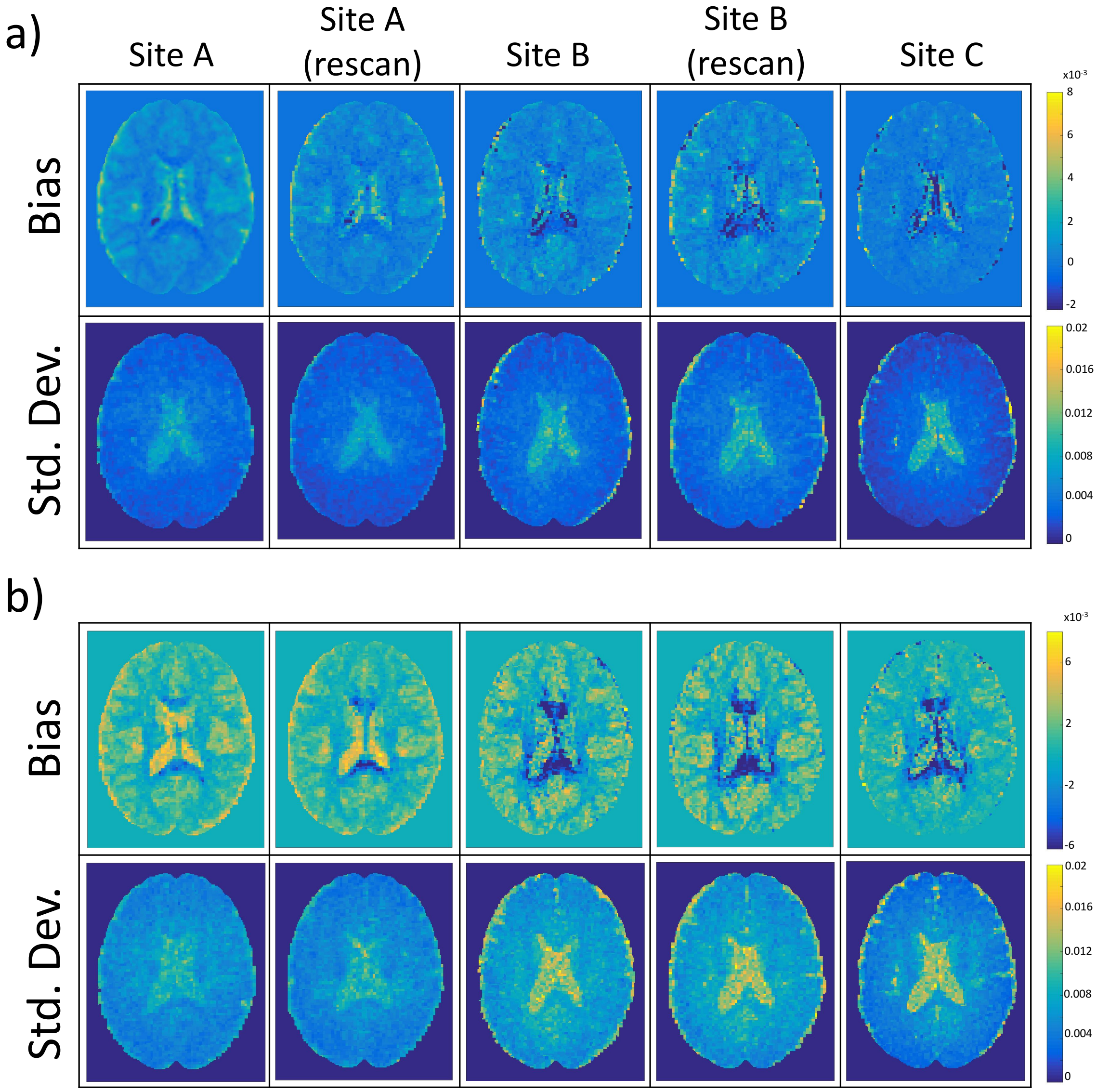

Figure 3.3: Spatial maps for the bias of GFA and the standard deviation of GFA for subject 01 are shown across all sites for (a) b=1000 and (b) b=2500. With these maps, the spatially dependent nature of the bias and variance estimates is evident. We can also see the variation in both estimates across the different scans. All images are shown on the scale of the unitless GFA measure.

of the ROI-averaged data (Figure 3.4) and found that the bias measurements were reasonably close to a Gaussian distribution. The plots reveal slight departures from normality due to heavier tails, though the distributions maintain a level of symmetry. In particular, we see a larger negative bias in the internal capsule of a single scan as well as larger standard deviations for the splenium of the corpus callosum for several scans. Note that the standard deviation of GFA is not expected to have a normal distribution, thus the departures seen in the Q-Q plot (pane c) are not worrisome.

These plots should be used as a visual check of the data for any major departures from the majority. A look into the raw data revealed that faulty segmentation was at fault and should be fixed for future analyses. Non-parametric methods may be useful for analyses with these slight departures from normality coupled with small sample sizes, though they are not required.

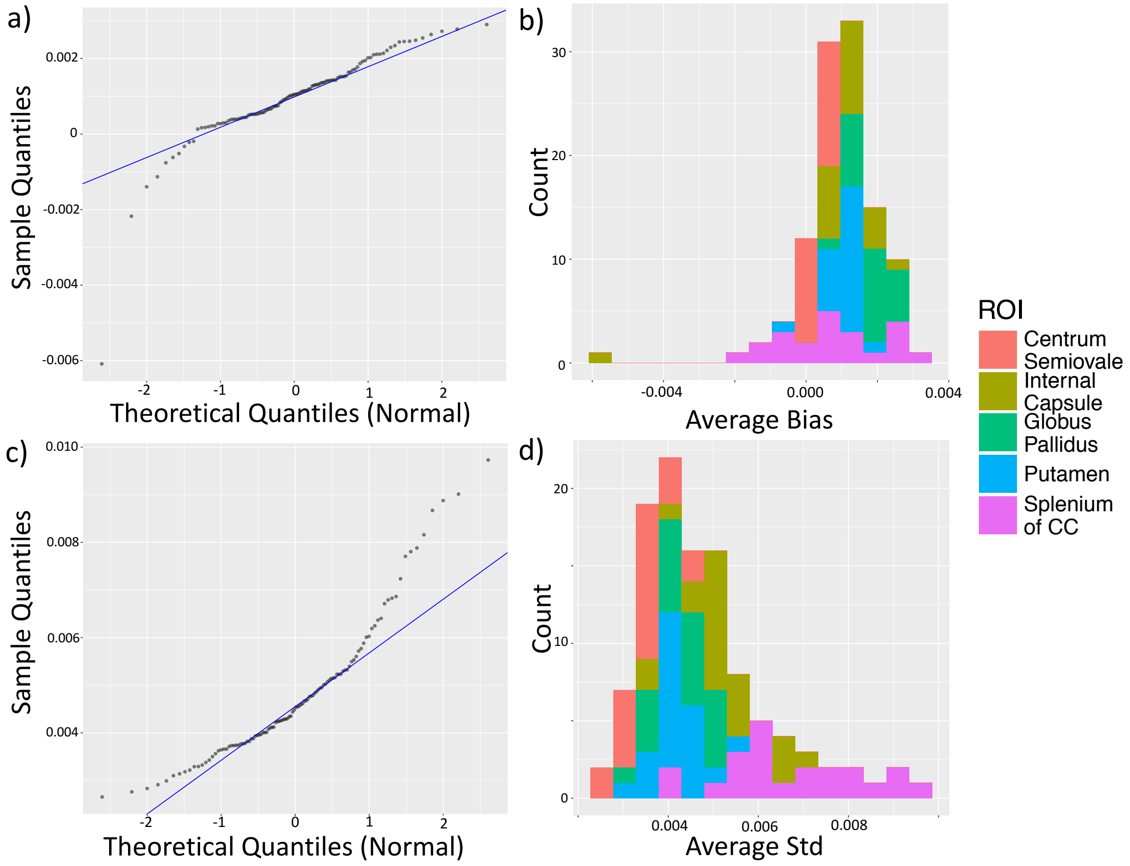

Figure 3.4: Checking for normality. For b=1000, Q-Q plot for bias of GFA (a) and histogram for bias of GFA (b) as well as the Q-Q plot for the standard deviation of GFA (c) and the histogram of the standard deviation of GFA (d) reveal distributions with deviations from the normal distributions, though normality is not expected for the standard deviation. The bias histogram reveals an outlier in the internal capsule, while the splenium of the corpus callosum appears to have higher than expected standard deviation. These deviations from expectation prompt a deeper look into the images to determine if there are issues in the data, such as artifacts or faulty segmentation.

We then created a scatterplot with the bias across the x-axis and the variance across the y-axis to assess the spread of the data and identify outliers (Figure 3.5). This scatterplot shows deviations of subject 00 at Scanner A from the other data points in terms of average bias for b=1000 s/mm^2 as well as several subjects and sites for b=2500 s/mm^2. In addition, two scans from Site B show larger than expected standard deviations. These outlying data points should be examined thoroughly before continuing the analysis.

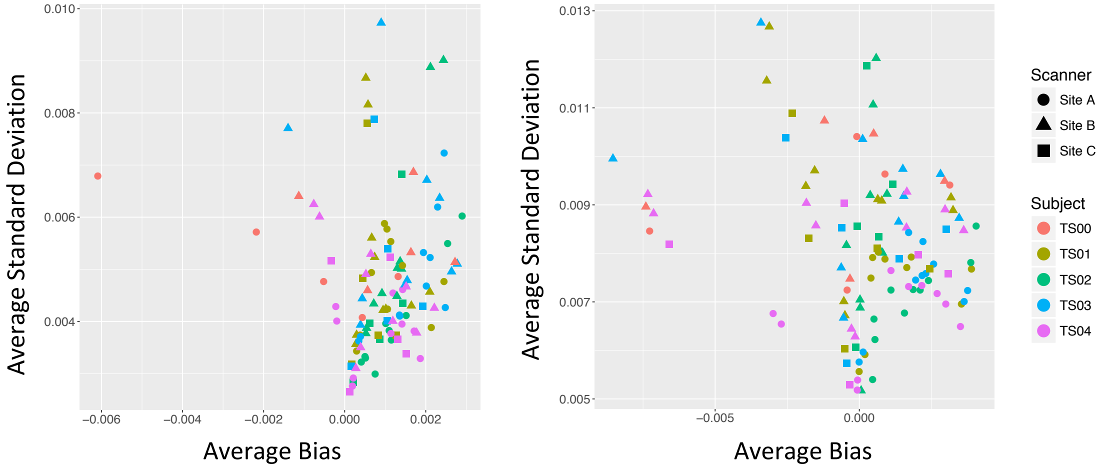

Figure 3.5: Scatterplots of average bias and average standard deviation for b=1000 s/mm^2 (left) and b=2500 s/mm^2 (right). Each data point is the average value within an ROI for each subject/scanner pair. In this plot, it is clear that one ROI from subject 00 at Site A has a larger negative bias than the rest of the scans at b=1000 s/mm^2. We also find several ROIs from a variety of subjects and sites have larger negative biases than the rest of the scans at b=2500 s/mm^2, as well as two scans from Site B which have larger than expected standard deviations. These data should be examined to ensure that these values are correct and that the quality is compatible with the remaining scans in this set. All bias and standard deviation values are shown on the scale of the unitless GFA measure.

Finally, we made boxplots for both bias and standard deviation of GFA. Figure 3.6 shows these boxplots for both b=1000 s/mm^2 and b=2500 s/mm^2. Each data point represents an averaged value across a single ROI for one subject at one site. This figure helps reveal patterns in bias and variance that may be due to either subject-specific variation or differences in acquisition protocols. According to our proposed rule of thumb, subject 00 at site A has larger than average bias in the internal capsule (b=1000 s/mm^2). For b=2500 s/mm^2, several scans appear to be outliers in terms of bias in the internal capsule. These results prompt the inclusion of a random effect for subject.

3.4 Discussion

This chapter demonstrates the importance of a review of scan quality when combining data from several study sites. With our method of scanning the same subject at multiple scan sites, we are able to estimate biases and variances that can be attributed only to differences in scan conditions, rather than subject-to-subject heterogeneity. The methods described in this chapter may be used to analyze the data quality of

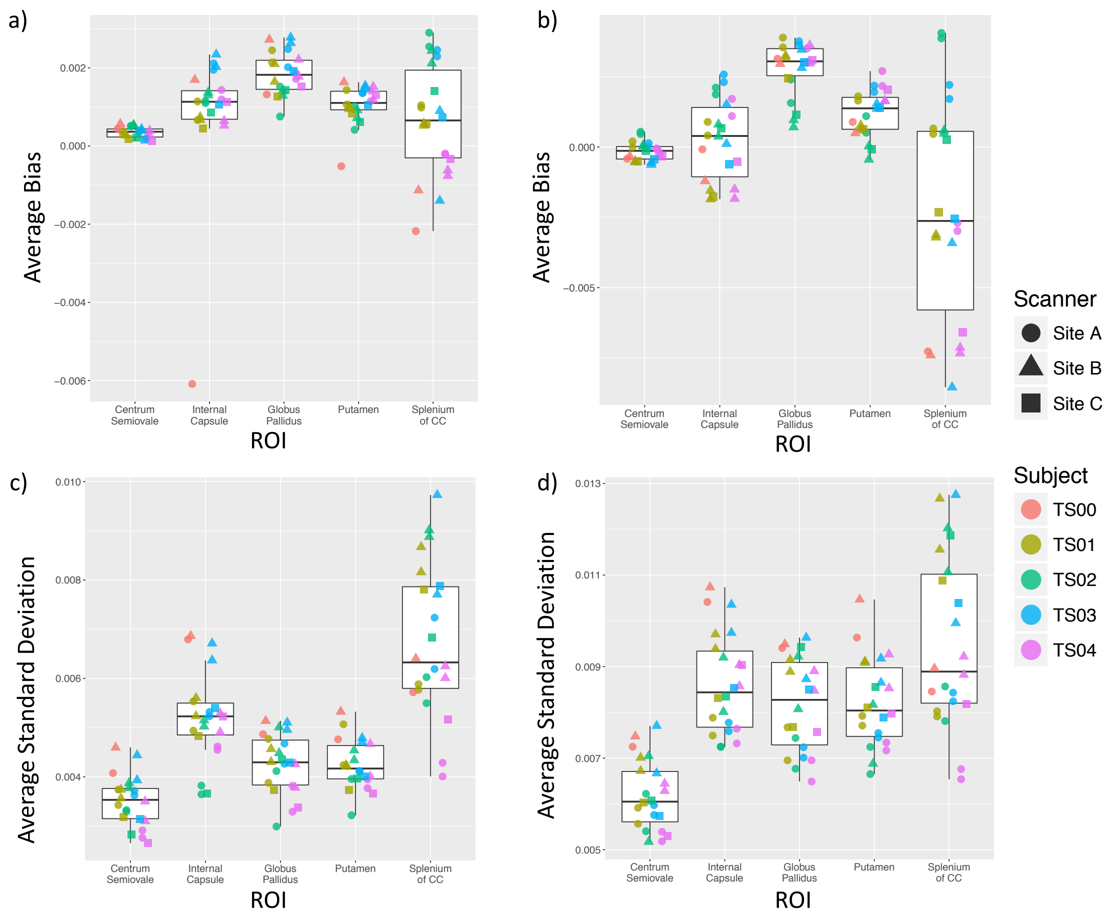

Figure 3.6: Additional exploratory box plots for (a) average bias for b=1000 s/mm^2 (b) average bias for b=2500 s/mm^2 (c) average standard deviation for b=1000 s/mm^2 (d) average standard deviation for b=2500 s/mm^2. Each point represents the value averaged across all voxels within the ROI. Scan sites are identified by the shape of the data point and each subject is identified by a different color. The whiskers of the boxplot represent 1.5 times the inter-quartile range (IQR). Any points falling outside this range are considered outliers and should be investigated further. We find that subject 00 had a larger than expected negative bias value in the internal capsule for b=1000 s/mm^2 which was the result of a segmentation error. In addition, several scans fell outside the IQR for the internal capsule on the b=2500 s/mm^2 scans. Each scan identified as an outlier should be examined for artifacts or other errors that could impact the analysis and corrected before continuing. All bias and standard deviation values are shown on the scale of the unitless GFA measure.

any study where heterogeneity between sites, scanners, or subjects is a concern. The decision tree in Figure 3.2 may be used to help guide the model selection process for analyses that combine data from different sources or where there is thought to be differences in the quality between different acquisitions.

CHAPTER 4

A DEEP LEARNING APPROACH TO ESTIMATION OF BIAS AND VARIANCE IN HIGH ANGULAR RESOLUTION DIFFUSION IMAGING

4.1 Introduction

Diffusion-weighted magnetic resonance imaging (DW-MRI) harnesses the diffusion of water for use as a proxy for underlying tissue microstructure. Diffusion tensor imaging (DTI) characterizes this microstructure, but cannot discern fibers in more than one direction per voxel, while high angular resolution diffusion imaging (HARDI) is able to discern crossing fibers. Harmonization of DW-MRI acquisitions remains an important, yet largely misunderstood area of re-search. DTI-derived metrics are known to have bias as a result of imaging noise (Basser (1997); Bastin et al. (1998); Skare et al. (2000); Basser and Pajevic (2000); Farrell et al. (2007); Hutchinson et al. (2017)).

Here we focus on a particular HARDI metric, generalized fractional anisotropy (GFA). The previous chapter provided methods for both bias correction and variance estimation of GFA from a single, empirical HARDI scan (Hainline, Nath, Parvathaneni, Blaber, Schilling, Anderson, Kang and Landman (2018)). Simulation Extrapolation (SIMEX) was used to estimate bias of GFA, while a wild bootstrap technique was used to estimate the standard deviation of GFA. While these methods work well, they tend to be computationally intensive due to the Monte Carlo simulations required for each estimate. In this work, we demonstrate a deep neural network approach for learning GFA itself, in addition to the bias and variance of GFA from the ob-served data. We find that we can take observed data, put it into our network and estimate a GFA value that is closer to the truth than what would result from calculating GFA from the diffusion orientation distribution function (ODF) of the original data via a regularized Q-ball fit.

Figure 4.1 maps out the relationship between the traditional statistical modeling techniques (SIMEX and wild bootstrap) and the deep learning approximations proposed herein. The former are relatively computationally intensive, while the latter are much faster to apply to new datasets, though they involve extensive training initially.

This work appears in Hainline, Nath, Parvathaneni, Schilling, Blaber, Anderson, Kang and Landman (2018).

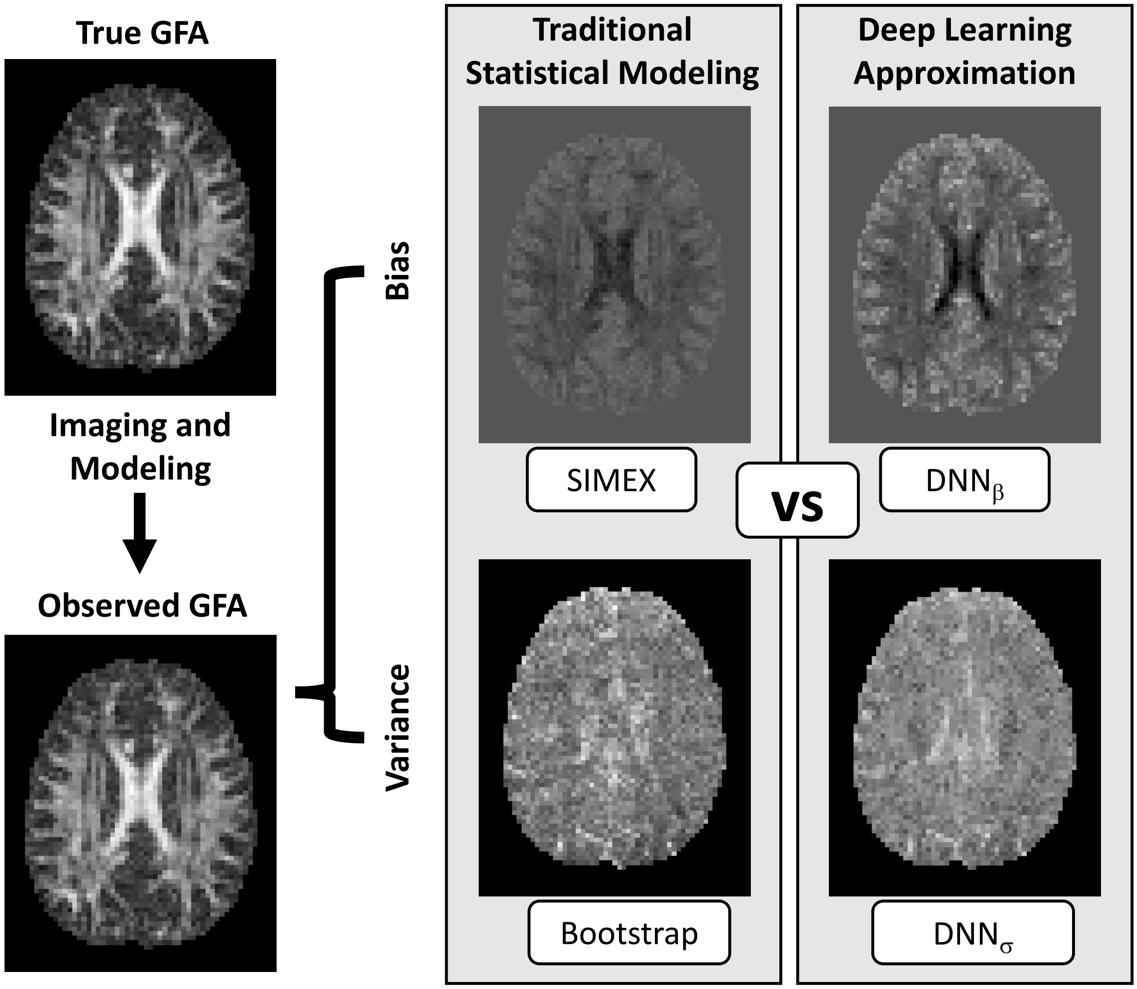

Figure 4.1: Broad overview of methodology presented. The true GFA is obscured by noise and artifacts in the imaging process. We previously used traditional statistical modeling to recover the true values through SIMEX and the wild bootstrap. In this chapter, we extend this idea and replace the statistical modeling techniques with deep learning approximations for bias and variance.

4.2 Methods

4.2.1 Data acquisition and preprocessing

The empirical data used in this experiment were obtained from a healthy volunteer 3T Phillips Scanner with a 32-channel head coil after informed consent (Nath et al. (2017, 2018)). The session consisted of 96 gradient directions at a b-value of 3000 s/mm^2. The voxel resolution is 2.5mm x 2.5mm x 2.5mm with 38 slices. The scan parameters were: Multi-Band=2; SENSE=2.2; TR= 2650 ms; TE=94 ms; partial Fourier=0.7. Fold over direction was A-P with a P fat shift. For each shell, an additional diffusion scan was acquired with reverse phase encoded volumes (i.e., fold over direction A-P with A fat shift) with a minimally weighted volume and 3 diffusion weighting directions with a b-value of 1000 s/mm^2 along the imaging frame cardinal directions, and all other parameters were kept constant.

A truth model is considered as the 'ground truth' data in this experiment. This model is generated from the spherical harmonic coefficients of the DW signal. This truth model is assumed noiseless and used as the basis for comparison of the methods.

To represent data from a typical DW acquisition, random Rician noise was added in quadrature to the ground truth data. The resulting 'observed' dataset, X_{obs}, represents an empirically observed HARDI acquisition, as the noise value is the standard deviation of the residuals, σ_E. The value of σ_E impacts the signal-to-noise ratio (SNR) of the observed data.

4.2.2 Preparation for analysis: calculation of true GFA, bias, variance

4.2.2.1 Calculating true GFA

The true GFA value is calculated via a regularized Q-ball imaging fit (Descoteaux et al. (2007)) to the true data model. For this work, a spherical harmonic basis was used in the reconstruction of the fiber orientation distribution function (ODF). GFA is given by

$$GFA = \frac{std(\psi)}{rms(\psi)} = \sqrt{\frac{m\sum_{i=1}^{m}(\psi_i - \overline{\psi})^2}{(m-1)\sum_{i=1}^{m}\psi_i^2}}$$

where ψ is the ODF vector, and $\overline{\psi}$ is its mean (Tuch (2004)).

4.2.2.2 Calculating true bias

We used a Monte Carlo approach to determine the true bias of an observed GFA value. This method involves simulating an observed voxel, calculating the GFA value for that voxel, subtracting the true GFA to determine the error for that observed voxel. This process is then repeated 100 times, after which the 100 errors are averaged, resulting in the true, voxel-wise bias of GFA.

4.2.2.3 Calculating true standard deviation

A similar approach is used for determining the true standard deviation of an observed GFA value per voxel. We simply take the 100 simulated GFA values from the same Monte Carlo procedure used for calculating true bias. Taking the standard deviation of these observed GFA values gives the true, voxel-wise standard deviation of GFA.

4.2.2.4 SIMEX bias estimation

The SIMEX was developed by Cook and Stefanski (1994) to correct for measurement error induced bias. In statistics, it is often used for the analysis of electronic medical records to correct human errors, but here we apply it to help correct errors induced by the imaging machinery itself. This method applies as long as the metric of interest changes monotonically as a function of noise and the noise distribution can be estimated (Cook and Stefanski (1994)). SIMEX utilizes the relationship between the noise level and the metric to estimate the potential, noise- (or error-) free value of the metric of interest. The method is simple and worked sufficiently well to quantify the bias and variance of GFA in an empirical HARDI acquisition. This method was first applied to DTI acquisitions (Lauzon et al. (2011)) and was recently expanded to apply to HARDI acquisitions in Hainline, Nath, Parvathaneni, Blaber, Schilling, Anderson, Kang and Landman (2018).

A brief explanation of the SIMEX procedure follows and full details of the algorithm can be found in Chapter 2 (Hainline, Nath, Parvathaneni, Blaber, Schilling, Anderson, Kang and Landman (2018)). SIMEX is built upon the idea that our observed data are a function of the true underlying data and random noise. For our application, all calculations are done per voxel, and the observed data are the result of adding stacked Rician noise with standard deviation σ_E to our noiseless truth data (Lauzon et al. (2013)):

$$\mathbf{X}_{obs} = \mathbf{X}_{truth} + \eta_{\sigma_E}$$

In order to estimate the bias of our metric, we must discern the relationship between the metric and the noise level. Thus, we generate data values with increasing amounts of noise (indexed by ω), calculate the metric, and observe the relationship between the ω and the metric. Once enough data points have been generated to establish a pattern, we can fit a quadratic curve and find our bias-corrected value by extrapolating backward. Note that other fits may be appropriate (i.e. linear or cubic), but we have found the best and most consistent results using a quadratic fit. The value of the noiseless metric can be found by extrapolating this curve to the point where $\omega = -1$, i.e. the point where there is no imaging noise.

Once we calculate the SIMEX-extrapolated GFA value, we can estimate the error of GFA by subtracting it from the observed GFA value:

$$\widehat{Bias} = GFA_{obs} - GFA_{SIMEX}$$

4.2.2.5 Bootstrap estimation of variance

The previously used statistical method for estimating the variance of GFA in HARDI acquisitions was the wild bootstrap. A bootstrap method is ideal, as they do not require the use of several repeated data acquisitions to estimate variance. In particular, the wild bootstrap is chosen due to its ability to estimate variances even when the model has heteroscedastic errors, as is the case with DTI data (Basser et al. (1994*a*); Whitcher et al. (2008)).

The wild bootstrap procedure is a modified residual bootstrap, where the first step is to compute the residuals between the model and the observed data. The signs of the residuals are then flipped randomly and added back to the observed data, resulting in a new, bootstrapped data set. This step is then repeated n times and the GFA is calculated for each of the n bootstrapped acquisitions. The standard deviation is then taken across all n simulated datasets as an estimate of the true standard deviation (Jones (2008)). Please refer to Chapter 2 for full details on the wild bootstrap procedure for HARDI data acquisitions.

4.2.3 Data processing for deep network

The data require further processing for input to our deep learning networks. The inputs for the deep neural networks are 6^{th} order spherical harmonic (SH) coefficients for each voxel. These coefficients were calculated as described in (Descoteaux et al. (2007)). The input for a single voxel is thus a 28×1 vector. The outputs for the deep neural networks are the true values of GFA, bias of GFA, and standard deviation of GFA as defined in section 4.2.2. The outputs are thus single values for each of the 3 networks.

The HARDI truth data consists of a single brain volume with 75 slices. First, we separated the training data from the validation data. Training data were defined as the first 41 axial slices of the volume, and the validation data were defined as the remaining 34 slices. The training and validation data sets were handled separately from this point on. Note that the partition is anatomically distinct, leaving no spatial consistency between training and validation sets. Separation of the training and validation data sets in this fashion yields results that are the worst-case scenario for our models.

The next step was to remove any background voxels, as our data was not masked prior to separation of training and validation data. This was done by removing all voxels with no diffusion information. After removal of these background voxels, the data are ready for fitting.

4.2.4 Network design

We train three distinct fully connected deep neural networks to predict (1) GFA, (2) the bias of GFA and (3) the variance of GFA. Each of the three networks takes 6th order spherical harmonic (SH) coefficients as input. The GFA network uses the true GFA as the output, the bias network uses the true bias as the output, and the variance network uses the true standard deviation as the output. Each network uses a mean squared error loss and Adam optimizer (Kingma and Ba (2017)). Training data were voxels from the observed data, $\mathbf{X}_{obs}$. The first 41 slices of our observed data were used as training data and the remaining 34 slices were reserved as a validation set. For each network, 40,942 voxels were included for the network training and 27,798 voxels were included for network validation. For training, 5-fold cross-validation was used to assess performance with 20% validation in each set. Root mean squared error was used to evaluate network performance for each. The validation error was computed at the empirically chosen epoch where the testing error was the smallest.

4.2.4.1 GFA network

The output of the GFA network is the voxel-wise true GFA. This network consists of 5 fully connected layers with 1200, 400, 200, 100, 66 neurons, respectively, with a single output neuron. The first three layers use a 'ReLU' activation function. The first two layers are followed by a 30% dropout layer to help prevent overfitting. A batch size of 1,000 was used.

4.2.4.2 Bias network

The output of the bias network is the voxel-wise true bias of GFA. This network consists of 5 fully connected layers with 400, 300, 200, 100, 66 neurons, respectively, with a single output neuron. The first two layers use a 'ReLU' activation function. The activation function is not used in the later layers to avoid constraint to positive values. The first two layers are followed by a 30% dropout layer to help prevent overfitting. A batch size of 1,000 was used.

4.2.4.3 Variance network

The output of the variance network is the voxel-wise true standard deviation of GFA. This network consists of 5 fully connected layers with 1200, 400, 200, 100, 66 neurons, respectively, with a single output neuron. The first three layers use a 'ReLU' activation function. The first two layers are followed by a 30% dropout layer to help prevent overfitting. A batch size of 1,000 was used.

4.2.5 Statistical analysis

Statistical significance of the proposed method is determined via a comparison of the squared errors of each method. For the GFA network, a two-way ANOVA was fit for the squared errors of the observed GFA, SIMEX-corrected GFA, and the DNN-predicted GFA. Following a significant result for the ANOVA, pairwise two-sample t-tests were conducted to determine pairwise significance after a Bonferroni correction. For both the bias and variance networks, two-sample t-tests were conducted between the statistical techniques and their DNN counterparts. All tests were conducted at a 5% significance level.

4.3 Results

Full details of the algorithm used in this analysis are shown in Figure 4.2. The results of the neural networks are compared to the results from the single observation as well as the statistical approaches (SIMEX and wild bootstrap). All calculations for the traditional statistical approaches were performed in Matlab version R2016a (MATLAB (2016)) and with the Camino Diffusion MRI toolkit (Cook et al. (2006)). The deep learning networks were trained using Python version 3.6.4 (Python Software Foundation) Python Core Team (2015) and the Keras deep learning library (Chollet et al. (2015)). All code, data, and trained models are available here: www.nitrc.org/projects/masimatlab under "Deep learning bias and variance (HARDI)."

The deep learning results are compared to the observed values as well as the SIMEX-corrected values for both GFA and bias, while the deep learning variance estimation is only compared to our SIMEX estimate.

Figure 4.3 plots the true values for GFA, bias of GFA and standard deviation of GFA against the observed, estimated, and predicted values. The observed GFA plot (A.1) shows a small amount of error, though only referring to a single scan, not an average across all possible scans. We found that our deep learning approach more effectively approximated the true GFA in comparison to the observed GFA (RMSE 0.0078 vs. 0.0082, p<0.001). The deep learning approach approximates the SIMEX estimate of GFA very well (RMSE 0.0078 vs 0.0078, p=0.987). Finally, the SIMEX estimate showed statistically significant improvement over the observed

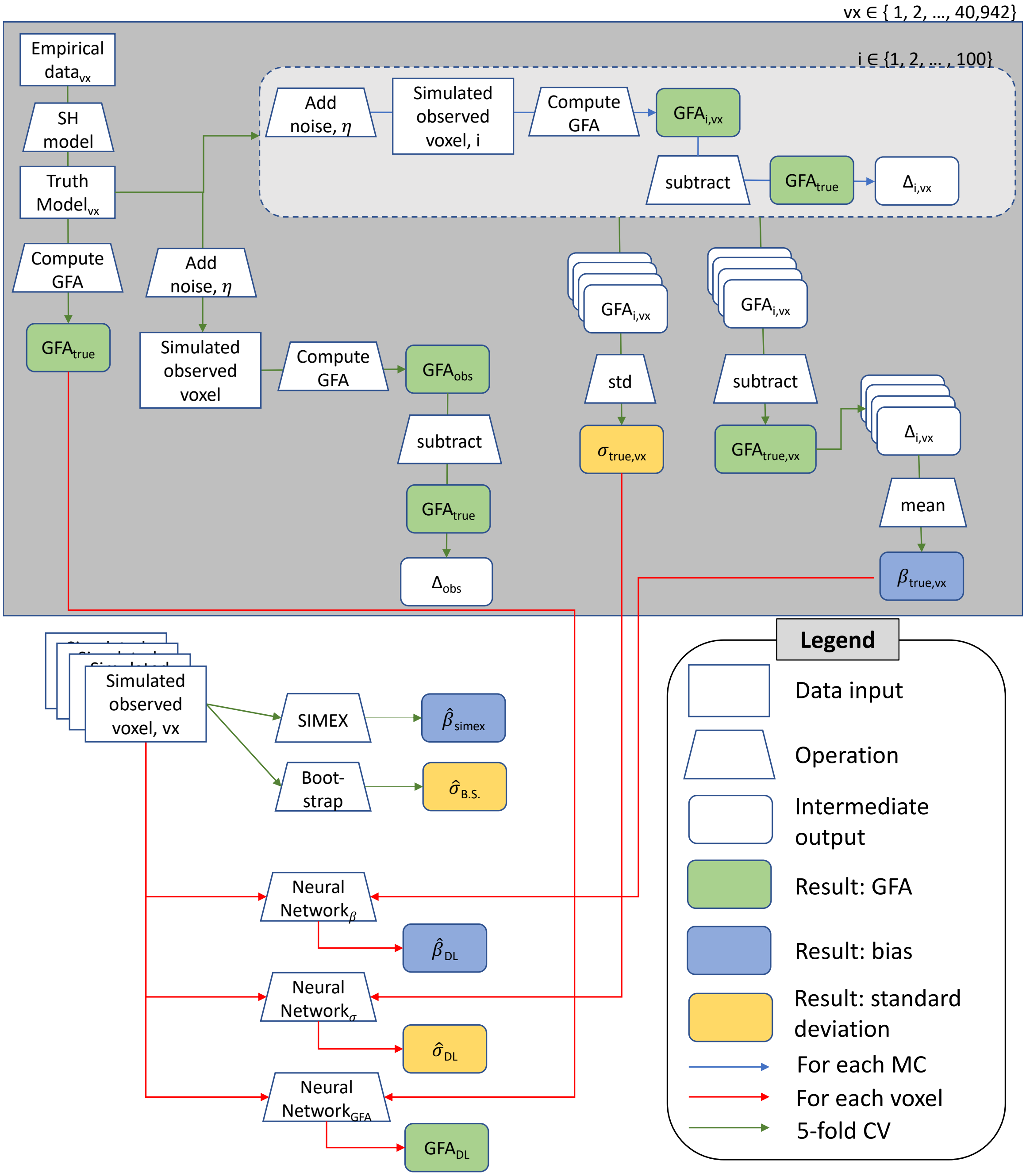

Figure 4.2: Overview of the algorithm used in this analysis. Data processing steps and statistical modeling procedures are located within the gray box. The results of the statistical procedures are com-pared to the deep learning approximations that are shown on the bottom half of the algorithm. In this algorithm outline, the shapes refer to data types, shape colors refer to result metrics and the arrow colors refer to the methodology employed.

estimate (RMSE 0.0078 vs. 0.0082, p<0.001).

The second column of Figure 4.3 refers to the performance of SIMEX and the deep learning network on the bias of GFA. The observed error of GFA is plotted against the true bias of GFA in B.1. We find that the neural network predicted bias is superior to the estimated bias of SIMEX (RMSE 0.0071 vs. 0.01, p<0.001).

The third column of Figure 4.3 demonstrates the performance of the wild bootstrap and the deep learning network on the standard deviation of GFA. The RMSE of the wild bootstrap is 12% lower than that of the deep neural network (RMSE 0.0011 vs. 0.00097, p<0.001).

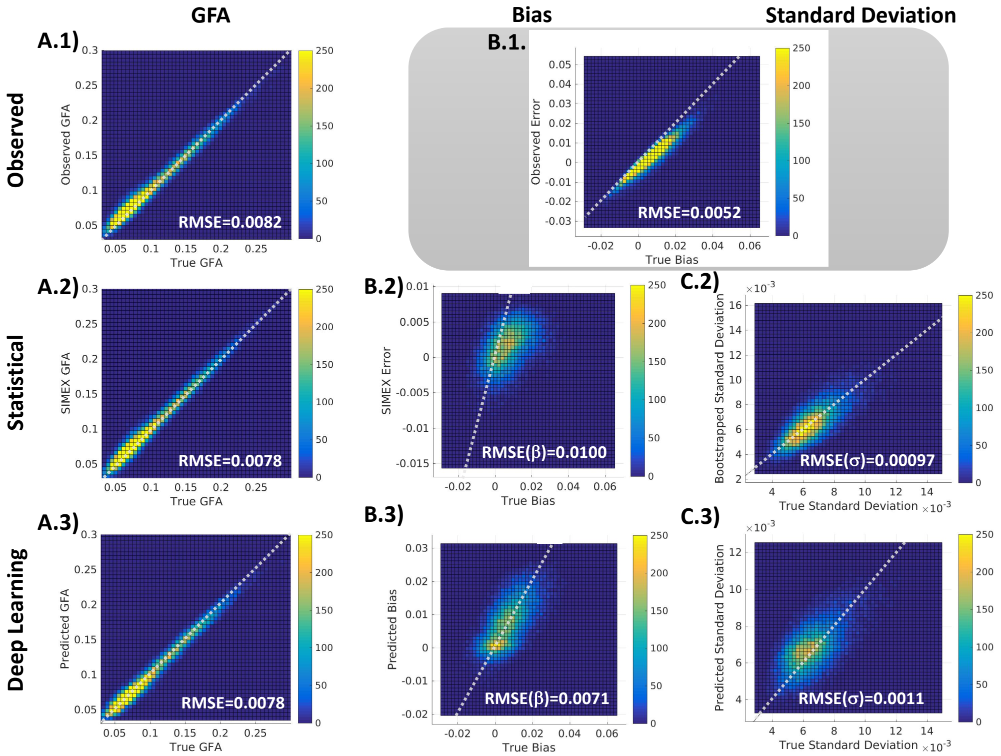

Figure 4.3: Quantitative results of the deep learning approximation in comparison to observed results as well as statistical results (SIMEX and bootstrap). The deep learning approach outperforms the Q-ball calculation of GFA from the observed data (A.1 and A.2) and has similar performance to the SIMEX-corrected GFA values (A.2, and A.3). The deep learning technique for bias prediction (B.3) results in a smaller RMSE in comparison to the SIMEX error estimation technique (B.2). The deep learning approach shows larger RMSE for standard deviation prediction (C.2) in com-parison to the wild bootstrap technique (C.1).

Figure 4.4 provides a qualitative look at the comparative performance of each method for GFA, bias of GFA and standard deviation of GFA. The true GFA, bias, and standard deviation are shown in the first row, with subsequent rows demonstrating the fitted values along with difference images. We find that our deep learning networks return appropriate values for GFA, bias, and variance in compliance with tissue microstructural differences. The gray box shows the error between the observed GFA and the true GFA as well as the difference between this ob-served error and the true bias of GFA.

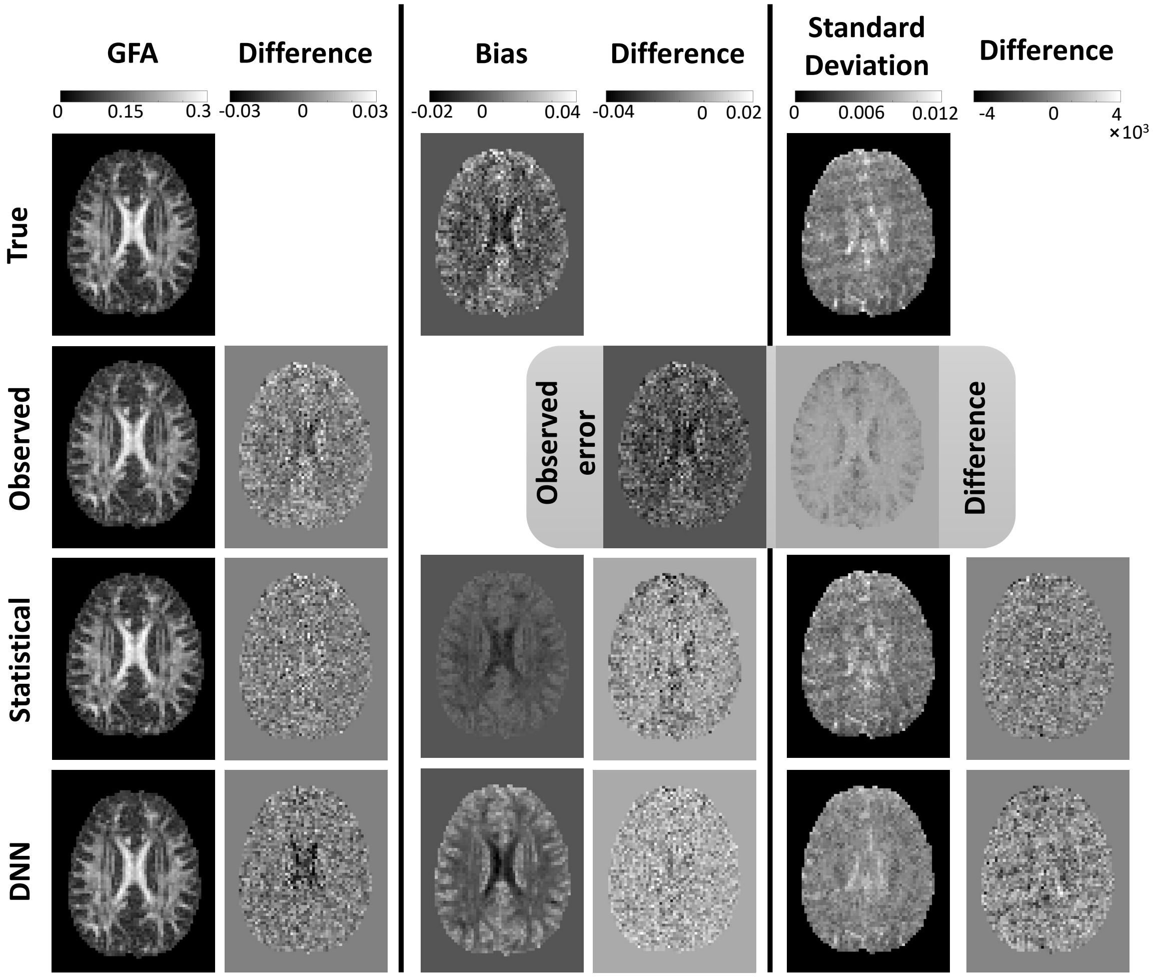

Figure 4.4: Qualitative results demonstrate the performance of the GFA, bias, and standard deviation estimation methods when compared to the ground truth values of each. We see that the DNN maintains the structural qualities of the brain and maintains comparable error when compared to the statistical techniques. Note that each column of images maintains the same color scale.

Figure 4.5 shows the training and cross-validation error curves for each of the three networks. The GFA network required 147 epochs of training and showed the least amount of overfitting. The bias network required much less training, reaching a minimum error before 100 epochs and demonstrated a significant amount of overfitting in later epochs. Finally, the standard deviation network trained very quickly and saw similar overfitting to the bias network as it trained for more epochs.

4.4 Discussion

Our previous work using SIMEX and the wild bootstrap to characterize HARDI data acquisitions have a variety of important applications to harmonization and data

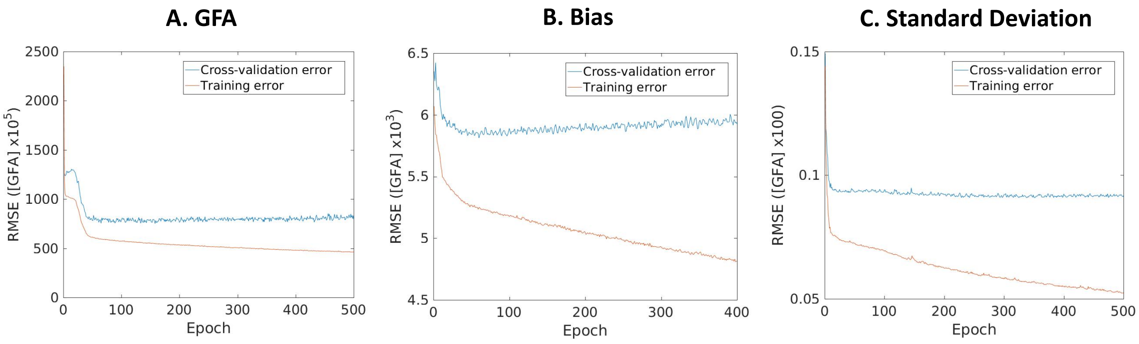

Figure 4.5: Training and cross-validation error curves for each of the three deep neural networks. The GFA network (A) shows no overfitting, while the other two networks demonstrate increasing overfitting with a larger number of epochs. All three networks train within 100 epochs. The final model is taken from the epoch with the lowest cross-validation error.

quality analyses. However, these methods can be time consuming and complicated for the casual user. The ability to evaluate and correct imaging metrics can allow for better inference and more replicable results. These methods can be especially useful in cases where brain changes are of interest, as the changes in brain microstructure are often very small and can be either diminished or magnified by bias.

In this chapter we have demonstrated the potential of a deep neural network to predict the true GFA value of a voxel more accurately than a regularized Q-ball fit on the observed data without a considerable increase in computation time. The SIMEX and wild bootstrap method can take up to ten hours per acquisition, where the trained neural networks provide results in 2-3 seconds. The neural networks explored herein are excellent candidates for inclusion in standard data processing procedures, as monitoring bias and variance of well-known metrics such as GFA is valuable for understanding data quality. While the networks demonstrated here are not perfect, these results are encouraging and reveal the potential of similar networks to enhance and support the traditional methods in diffusion imaging.

Future work will be to train networks for various noise levels and b-values, as this work was done with a single noise level at b=3000 s/mm^2. The ability to incorporate SNR information can only improve the performance and usefulness of these networks.

CHAPTER 5

EVIDENCE-BASED INFERENCE ON RESTING STATE FUNCTIONAL CONNECTIVITY

5.1 Introduction

Resting-state functional magnetic resonance imaging (rs-fMRI) techniques measure the blood oxygen level-dependent (BOLD) signal in the brain while at rest. Though it is traditionally believed that brain regions that have correlated activation patterns are likely part of the same functional network, it can be unclear whether temporally correlated signals are indeed functionally connected, or if that signal is the result of some combination of imaging noise and random chance (Logothetis and Wandell (2004)). In such cases where the signals are noisy, methodologies that can distinguish between these truly correlated signals and imaging artifacts are essential.

The two most common analysis methods for resting state fMRI data are correlation analysis and independent component analysis (ICA). This work focuses on correlation analysis, where the correlation coefficient is calculated for pairs of voxels or regions of interest (ROI) and that correlation is determined to be significant or not based on statistical inferential techniques (Lee et al. (2013)). Analyses often use pairwise t-tests in which t-statistics are calculated for each pair of voxels or ROIs and a significance threshold is determined for the entire brain. However, this threshold must be chosen very carefully in order to avoid an inflated family-wise Type I error (false positive) rate due to the large number of simultaneous comparisons. One approach for addressing the Type I error inflation is to instead focus on controlling the false discovery rate (FDR) (Benjamini and Hochberg (1995)). Controlling the FDR is a less conservative approach than a Bonferroni correction controlling family-wise Type I error rate and thus may be more favorable for use with many comparisons, such as brain activation studies. In addition to FDR control, another common method is to use a permutation testing technique, which utilizes resampling of the observed data to create a sampling distribution which is then used to determine statistical significance (Holmes et al. (1996); Nichols and Holmes (2001); Hayasaka and Nichols (2004); Nichols and Hayasaka (2003)). Both of these methods, however, can often result in an inflated Type II error (false negative) rate.

An inflated global Type I error can be prevented via the likelihood paradigm (Royall (1997); Blume (2002)), which minimizes the weighted average of false positive and false negative error rates, rather than fixing the Type I error and maximizing the power as in traditional frequentist hypothesis testing. Previous work by Kang et al. (2015) has shown promise in the likelihood paradigm in task-induced fMRI, and herein we venture to extend the procedure to an evaluation of resting-state functional connectivity.

This work appears in Hainline and Kang (2018).

5.2 Likelihood Paradigm

The evidential framework aims to explain what the data themselves say about the proposed hypotheses. The likelihood principle states that under a probability model, all of the evidence contained in the data is summarized in the likelihood function. Further, the Law of Likelihood (See Section 5.6) implies that the better supported of the pair of hypotheses is the one that assigns the higher probability to the observed data. This is measured via the likelihood function and the likelihood ratio (LR).

Let X be a random variable that follows the distribution $X \sim f(X;\theta)$ where θ is the parameter of interest. If we observe $\underline{X} = \underline{x}$, the likelihood function is given by $L(\theta|\underline{x})$. The likelihood function is the probability density function for a fixed parameter, θ, and is used after the data are observed.

Consider two hypotheses, $H_A : \theta = \theta_A$ and $H_B : \theta = \theta_B$. $P(\underline{x}|H_A)$ is the probability of observing $\underline{x}$ given that H_A is true, and $P(\underline{x}|H_B)$ is the probability of observing $\underline{x}$ given H_B is true. The ratio of these two probabilities is the likelihood ratio. The likelihood ratio measures the strength of the evidence supporting one hypothesis over another.

$$LR = \frac{P(\underline{x}|H_A)}{P(\underline{x}|H_B)} = \frac{L(\theta_A|\underline{x})}{L(\theta_B|\underline{x})}$$

Likelihood ratios require the explicit definition of two competing hypothesis because, by definition, evidence under the likelihood paradigm is relative. The LR can only show support for one hypothesis over another or show a neutral result when the data do not favor a single hypothesis. In summary, a $LR = 1$ is neutral evidence, a $LR > 1$ shows support for H_A over H_B, and a $LR < 1$ shows support for H_B over H_A.

A key result of the likelihood paradigm is the convergence of both global Type I and Type II error analogs to zero as the information in the sample, i.e., sample size, increases, whereas traditional methods may never reach a Type I error below the pre-specified size of the test. Since the LR is only the measures of the strength of evidence, rather than the size of the test, the LR does not need to be adjusted for simultaneous comparisons. The global error rate converges to zero rather than being inflated by simultaneous comparisons (Kang et al. (2015)).

Finally, the likelihood paradigm is unique in that it divides evidence into three distinct regions. These regions are defined by a parameter, k, which represents a guidepost for the definitions of strong and weak evidence. Further, the probability of observing misleading evidence is bounded above by the value $1/k$ (Royall (1997); Blume (2002)). A likelihood ratio less than $1/k$ falls in the strong evidence region supporting the null, while a likelihood ratio greater than k falls in the strong evidence region supporting the alternative. Any LR between $1/k$ and k falls in the weak evidence region, which corresponds to inconclusive evidence. This weak evidence region is what allows the likelihood paradigm to maintain lower error rates even under multiple simultaneous tests.

As the statistical information in the sample increases (i.e. increased sample size or time series length), the error rate in this region will shrink to zero and all LR will be classified in the strong evidence categories (Kang et al. (2015)).

For a more detailed discussion of the likelihood paradigm along with derivations see Section 5.6.

5.3 Methods

5.3.1 Data

5.3.1.1 Simulated data

To evaluate the performance of the proposed methodology in comparison to traditional methods, we have conducted a simulation study. The simulated data consisted of 6 ROIs in addition to an additional region to represent cerebrospinal fluid (CSF). Pairs of ROI were given true temporal correlation values of either 0 (to represent truly null ROI pairs) or a value between 0.45 and 0.57 (to represent truly functionally connected ROI pairs). The data were generated to be both spatially and temporally correlated. The spatial correlation was applied via an exponential covariance function with a unit decaying parameter. The data were simulated using an AR(1) temporal correlation with $\phi = 0.6$, which controls the correlation between observations separated in time. Data were generated across a range of time series lengths: T=64, 128, 256, 320 scans. 300 simulations repetitions were conducted for each combination of parameters.

Before analysis, spatial smoothing was performed with a Gaussian filter with $\sigma = 1.5$ as in Kang et al. (2015). Each ROI consisted of 100 voxels (10×10) and the time series were normalized and averaged across all 100 voxels for analysis.

5.3.1.2 Clinical data

For the real data application, we used a sample of 29 healthy volunteers between the ages of 20 and 50 years old. These subjects had no history of psychiatric disorders or psychotropic medication use. After informed consent, each participant was scanned on a Siemens 3.0 Tesla Trio Tim scanner with an 8 channel head coil. Each received a T1-weighted 3D Magnetization-Prepared Rapid Gradient-Echo (MPRAGE) sequence with a repetition time of 2300 ms, echo time of 3.46 ms, a flip angle of 9 degrees with a voxel size of 0.9 x 0.9 x 1.2 mm as well as an Echoplanar Blood Oxygen Level Dependent (EpiBOLD) functional resting-state scan with repetition time of 2000 ms, echo time of 27 ms with a voxel size of 4.0 x 4.0 x 4.0 mm.

Preprocessing included slice timing correction, head motion correction across all scans, co-registration, and normalization to the standard Montreal Neurological Institute (MNI) template. All preprocessing was performed using the FSL software package (Smith et al. (2004)). Additionally, the scans were segmented into a CSF region and 14 ROIs chosen from the Default Mode Network (DMN) and defined using Automatic Anatomical Labeling (AAL) (Tzourio-Mazoyer et al. (2002); Raichle et al. (2001)).

5.3.2 Null distribution

This methodology relies on the definition of a null distribution to define the alternative hypothesis for the Likelihood technique. Here we operate under the assumption that any correlation that an ROI has with the CSF region is the result of noise, rather than actual functional connectivity (FC) signal. The correlations present between the CSF region and the ROIs is assumed to consist of all noise, either physiological or random, that has not been fully taken care of in the preprocessing of the data.

We create the null CSF distribution by computing the correlation coefficient between the average time series signal in each ROI and each voxel within the CSF region. These correlations form the null CSF distribution that will be used to define the alternative hypothesis.

5.3.3 Hypotheses

Since the likelihood approach defines evidence to be relative, we must have two well-defined hypotheses for comparison. A value must be chosen to represent the cutoff between null and functionally connected ROI and serve as a basis for the null and alternative hypotheses. fMRI literature supports 0.3 as an acceptable value that is large enough to not be the result of physiological or imaging artifacts (Cordes et al. (2002)). Thus, we have chosen to use 0.3 as the correlation coefficient value for our null hypothesis.

For the alternative hypothesis, we must choose a value that is able to consistently differentiate between unconnected and functionally connected ROI pairs. We use the null CSF distribution described in Section 5.3.2 as a guide for determining this value. Since this distribution is a good indicator of the noise in the data, we use the interquartile range (IQR) of the null distribution to help define the alternative hypothesis as follows:

$$\begin{aligned} H_0 &: FC_{j,k} = 0.3 \\ H_1 &: FC_{j,k} = 0.3 + 3 \times IQR \end{aligned}$$

where IQR is the interquartile range of the null CSF distribution. The use of the null CSF distribution in defining the alternative hypothesis allows for the hypothesis to change in relation to the quality of the data. Noisy data will result in a wide null CSF distribution, leading our alternative hypothesis to be more extreme.

5.3.4 Statistical analysis

5.3.4.1 Simulated data

Performance is assessed in terms of Type I and Type II error rates for the proposed method as well as a traditional t-test controlling for a false discovery rate of 5% and a permutation test.

For the likelihood technique, the likelihood ratio, $L(H_1)/L(H_0)$, was used to determine if the observed correlation between each pair of ROIs showed strong evidence for the null, strong evidence for the alternative, or inconclusive evidence. We chose to use seta cutoff of $k = 20$, which controls the probability of observing misleading evidence at 1/20 (0.05) (Royall (1997); Blume (2002)). Thus, a $LR < 1/20$ indicated strong evidence supporting the null, $1/20 < LR < 20$ indicated the inconclusive (or weak) evidence region, and $LR > 20$ indicated strong evidence supporting the alternative.

The t-test results were controlled at a false discovery rate of 5% in an effort to maintain a fair comparison between methods. In addition, a permutation test was applied to the correlations to test if they differed from zero.

In addition, the same analysis was also conducted using the dichotomous likelihood paradigm (DLP), which better mimics the accepted statistical hypothesis testing techniques in that it eliminates the weak evidence region in favor of two regions: evidence in favor of the null hypothesis and evidence in favor of the alternative hypothesis. Under the DLP, we adopt $k = 20$, so any $LR > 20$ is considered evidence supporting the alternative hypothesis and any $LR < 20$ indicates evidence supporting the null hypothesis.

5.3.4.2 Clinical application

A data decimation approach is used to assess the performance of the proposed method on clinical data. According to each testing strategy, each ROI pair was categorized as 'significant' or 'not significant' in terms of functional connectivity using the full sample of 29 subjects. These results are considered the "truth" for the decimation procedure. The analysis is repeated for increasingly reduced sample sizes, and, for each sample size, the results with the smaller sample are compared to the "truth" results in order to determine error rates. In this procedure, the ability to reproduce the "true" results is tested.

While this method relies on the accuracy of the initial full sample results and may not give any indication to the true error rates, it does allow for the examination of each strategy's behavior as the sample size decreases. Generally, a method that can maintain the results of a larger sample with a smaller sample is more favorable.

5.4 Results

5.4.1 Simulation study

The results of the simulation study are shown in Figure 5.1 and Table 5.1. The t-test and the permutation test show very similar behavior, as expected. We find that, across the range of sample sizes and time series lengths, the likelihood method maintains a much lower false positive rate than the remaining two methods (first column). For example, at the moderate sample size of 30 subjects, the likelihood method results in a 69% decrease in false positive rate compared to the t-test for a time series length of 64 seconds and a 98% decrease for T=320 seconds. The results are very similar when comparing the likelihood method to the permutation test.

Note that the false positive rates for both the FDR-corrected t-test and the permutation test are much higher than expected. This is due to the ROIs being averaged

Table 5.1: Average error rates for the T-test, Permutation test, likelihood paradigm, and dichotomous likelihood paradigm approaches in simulation. The results are presented as functions of both time series length (T) and sample size(N).

T	N	T-test	Permutation test	LP	DLP
64	10	0.339	0.343	0.203	0.312
	20	0.368	0.374	0.236	0.297
	30	0.382	0.387	0.247	0.286
	40	0.388	0.393	0.254	0.281
	50	0.394	0.398	0.255	0.282
128	10	0.313	0.321	0.100	0.189
	20	0.347	0.355	0.117	0.167
	30	0.364	0.369	0.126	0.161
	40	0.375	0.378	0.130	0.159
	50	0.378	0.381	0.134	0.158
256	10	0.269	0.277	0.021	0.066
	20	0.303	0.310	0.023	0.044
	30	0.321	0.326	0.026	0.039
	40	0.332	0.338	0.026	0.038
	50	0.340	0.344	0.026	0.034
320	10	0.261	0.268	0.006	0.031
	20	0.299	0.305	0.006	0.017
	30	0.316	0.323	0.006	0.010
	40	0.330	0.335	0.006	0.009
	50	0.338	0.346	0.005	0.007

across the individual voxels. While the time series for individual voxels follows an AR(1) temporal correlation process, the averaging tends to result in systematic bias. The correlations resulting from these averaged values will exhibit both biased point estimates and variances. The t-test and permutation test rely on assumptions regarding the distribution of the correlations, and with an averaged time series, these assumptions are not upheld. The likelihood technique, however, takes these biases into account via the CSF null distribution, resulting in much lower false positive rates.

We also see that the likelihood method results in higher false negative rates for T=64 and T=128 (across all n), though this rate drops dramatically for T=256 and T=320 (second column). The averages of the false positive and false negative rates are shown in the third column, where the superiority of the likelihood method is quite clear. The T-test and permutation tests maintain average error rates greater than 20% across all sample sizes and time series lengths, while the likelihood method shows dramatic decreases in the average error as the time series length increases.

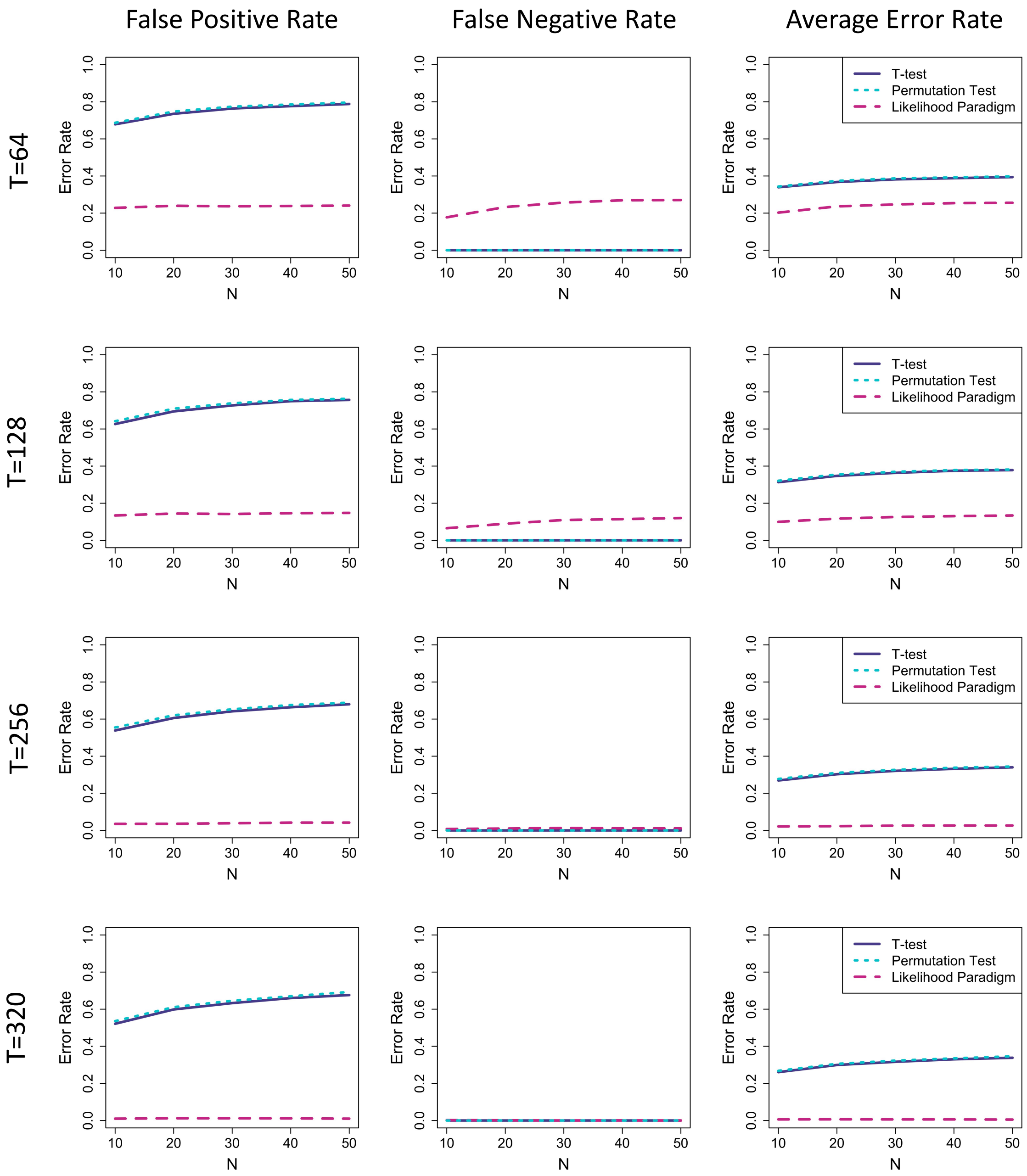

Figure 5.1: Results of the simulation study. Rows represent different time series lengths and columns represent the different errors. The first column shows the error rate among truly null ROIs (false positive). The second column shows the error rate among truly connected ROIs (false negative). The third column is the average of the first two columns. The LP approach maintains a much smaller false positive rate than the traditional methods (which show very similar behavior across all simulations. For T=64 and T=128, we see an increased false negative rate for the LP approach, where the other two have extremely small errors. However, the LP technique demonstrates a lower average error for all simulation settings. Note that the errors remain relatively constant across sample sizes and decrease substantially as the time series length increases.

Specifically, at n=30, the likelihood method shows a 35% lower error than both the t-test and the permutation test at T=64 seconds, which rises to 98% lower at T=320 seconds. The average errors for the likelihood method drop to 0.026 at T=256 seconds and 0.005 at T=320 seconds.

Further, Figure 5.2 demonstrates the error rate that occurred in the inconclusive region for the simulation study. The number of simulations that result in an inconclusive result drop as a function of sample size for all time series lengths. The longer the time series, the more information present in the data, leading to smaller errors and a faster convergence to zero than the shorter lengths, though even the shorter time series length show a decreasing pattern.

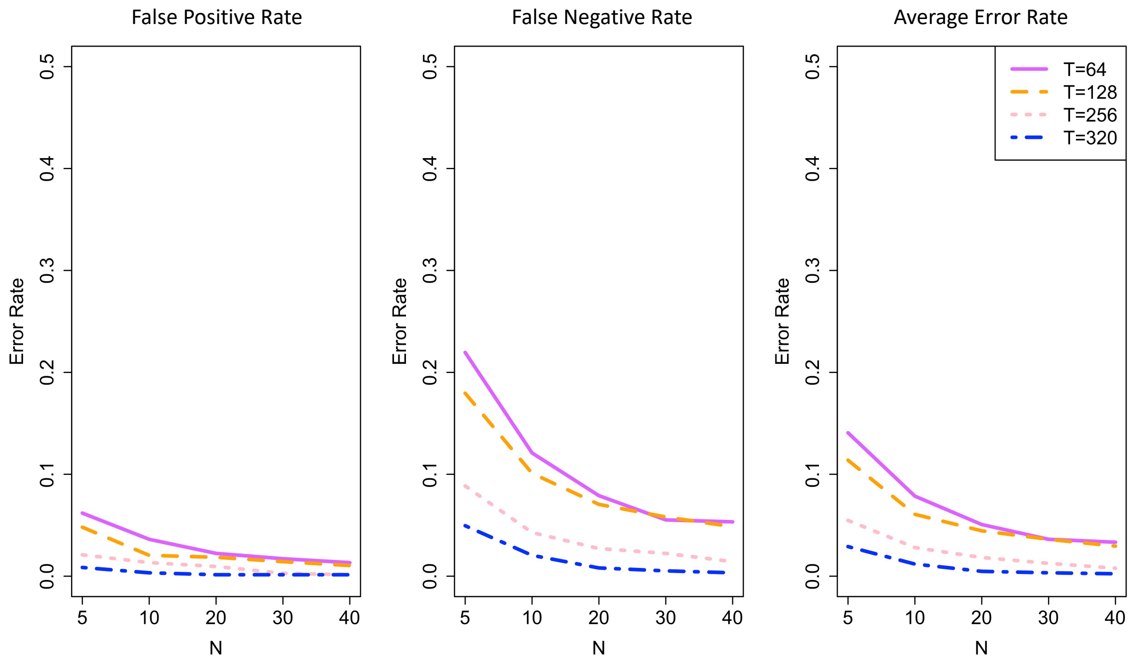

Figure 5.2: Error rates within the inconclusive region. The error decreases uniformly as the time series length increases and as the sample size increases. If the sample size were allowed to increase to infinity, the errors for each time series length would converge to zero.

Results from the DLP approach are very similar to those of the LP and are shown in Section 5.8.

5.4.2 Clinical application

The average error rates of each of the three methods for data decimation procedure are shown in Figure 5.3. The proposed likelihood method demonstrates increased robustness against reduced sample size than both the traditional T-test and the permutation test. The likelihood method maintains the results of the full sample until the sample size decreases to n=11, when the false negative rate begins to rise. This is due to a larger number of simulations being deemed inconclusive. Average error rates for the t-test rise from 12% (n=26) to 34% (n=5). The permutation test shows slightly higher average error rates, from 13% (n=26) to 50% (n=5).

For the t-test and permutation test we see a decrease in false positives and an increase in false negatives as the sample size decreases. We see slight increases in both error rates for the likelihood method, but neither error rises above 20% for any sample size.

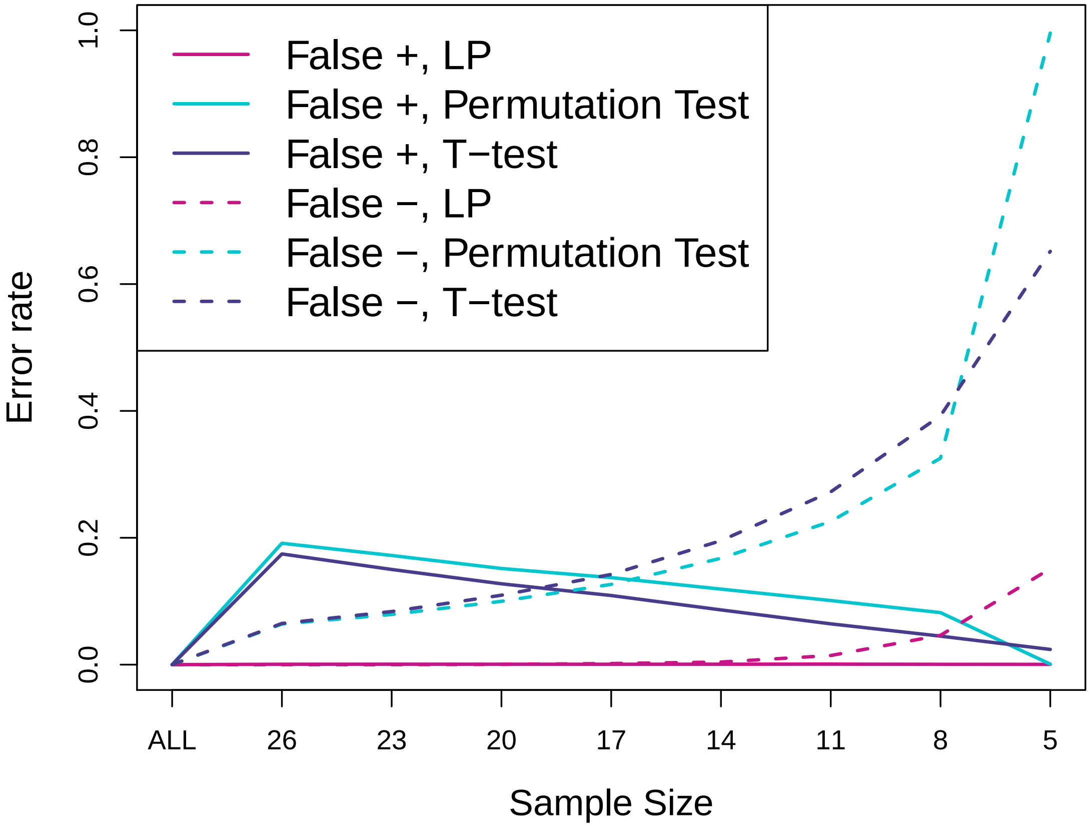

Figure 5.3: Data decimation results. The likelihood paradigm approach maintains very low Type I and Type II error rates down to a sample size of 8 subjects. The t-test and permutation test results, however, show higher Type I and Type II error rates across all sample sizes. As the sample size decreases, the frequentist approaches become less likely to reject a truly null pair of ROIs and more likely to fail to reject a truly connected pair.

5.5 Discussion

Brain functional connectivity studies require a large number of simultaneous comparisons in order to determine functionally connected pairs of brain regions. Due to these multiple comparisons, it is extremely important to choose statistical methods that are able to control Type I and Type II errors under these conditions. In both a simulation study and a data decimation example using clinical data, the LP approach proposed herein has shown better behavior in terms of these errors than both an FDR-corrected t-test and a permutation test. The likelihood method resulted in up to a 98% decrease in false positive rate and average error rates as low as 0.005, in simulation. Further, we have shown that the likelihood methods are more robust to decreases in sample size than the conventional approaches via a data decimation study. Thus, the use of the likelihood approach will allow for researchers to identify more regions of true functional connectivity while not risking a similar increase in false identifications.

Future work in this area will aim to incorporate spatial information in an effort to improve the results. Due to the spatially dependent nature of fMRI data, it is likely that taking into account any spatial correlation will allow for improved accuracy with a smaller sample due to the amount of information gained by using of neighboring voxel information to inform the cross-correlation of each. Further, we chose to average our voxel-wise information across ROI in order to get an ROI-level result. The next step for the likelihood method would be an application to voxel-level data. This will likely improve results, as well, as ROI-averaging tends to result in a loss of statistical power.

In addition, one of the strengths of the likelihood technique is the ability to choose the weighting of Type I and Type II errors via the choice of the null and alternative hypotheses. The hypothesis definitions provided herein should serve as guides, not absolutes. The definition of these hypotheses can be adjusted based on prior knowledge of the data behavior and noise patterns in order to allow for optimal performance depending on the needs of the specific analysis. Further, depending on the goals of the analysis, the researcher may wish to adjust the study design in order to control the amount of misleading evidence. For example, when determining active brain areas in preparation for surgery, an inconclusive result is of little use. The Likelihood method differs from hypothesis testing in that it can distinguish between three areas, those that are active, inactive, and inconclusive, whereas hypothesis testing can only distinguish between active areas and inconclusive areas. Thus, the solution when correct results are imperative would be to obtain a longer time series in order to shrink the weak evidence region and ensure that any activity can be detected. Likelihood methods provide a different approach than hypothesis testing, and must be tailored to the needs of the individual study in order to secure the highest benefit.

5.6 Review of the Law of Likelihood

The Law of Likelihood was first presented by Hacking (1965) and later popularized by Royall (1997):

The Law of Likelihood: If hypothesis A implies that the probability that a random variable X takes the value of x is $p_A(x)$, while hypothesis B implies that the probability of $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence.

To further demonstrate the law, consider the following example. We have a random variable X that follows a probability distribution with parameter θ, thus observation x provides a likelihood function, $L(\theta; x)$. Consider the simple null hypothesis $H_0 : \theta = \theta_0$ and simple alternative hypothesis $H_1 : \theta = \theta_1$. According to the Law of Likelihood, observation $X = x$ provides evidence supporting the alternative hypothesis if and only if $L(\theta_1; x) > L(\theta_0; x)$ and the ratio $L(\theta_1; x)/L(\theta_0; x)$ measures the strength of that evidence.

In short, the Law of Likelihood concludes that the hypothesis with the higher likelihood given the observed data is the better supported of the two.

5.7 Review of the Probability of Observing Misleading Evidence

Under the likelihood paradigm, misleading evidence is defined as the conclusion of strong evidence in favor of the wrong hypothesis over the correct hypothesis. The probability of observing misleading evidence is a property of the study design and does not apply to any specific set of observed data (Blume (2002)). For any fixed sample size, the probability of observing misleading evidence of strength k or greater is always bounded above by $1/k$.

Following the notation from Blume (2002), assume $f(X)$ and $g(X)$ are both probability density functions, and that $X \sim f(X)$, then

$$P_f\left(\frac{g(X)}{f(X)} \geq k\right) \leq \frac{1}{k}$$

Please refer to Blume (2002) and Kang et al. (2015) for a full discussion of this relationship and derivation of the asymptotic properties of the probability of misleading evidence.

5.8 Results of a Dichotomous Likelihood Paradigm (DLP) approach

This section contains the results of a dichotomous likelihood paradigm (DLP) approach to the same simulation study detailed in Section 5.3.4.1. We have provided these results in Figure 5.4.

Note that the false positive rates (first column) are identical to those from the original likelihood paradigm. This is because the DLP changes the way we conclude in favor of the null, but not the way we conclude in favor of the alternative. In both techniques, we conclude in favor of the alternative if the $LR > 20$. However, for the DLP, we collapse the weak evidence region with the strong evidence region in favor of the null, resulting in slightly different values for the second and third columns of Figure 5.4.

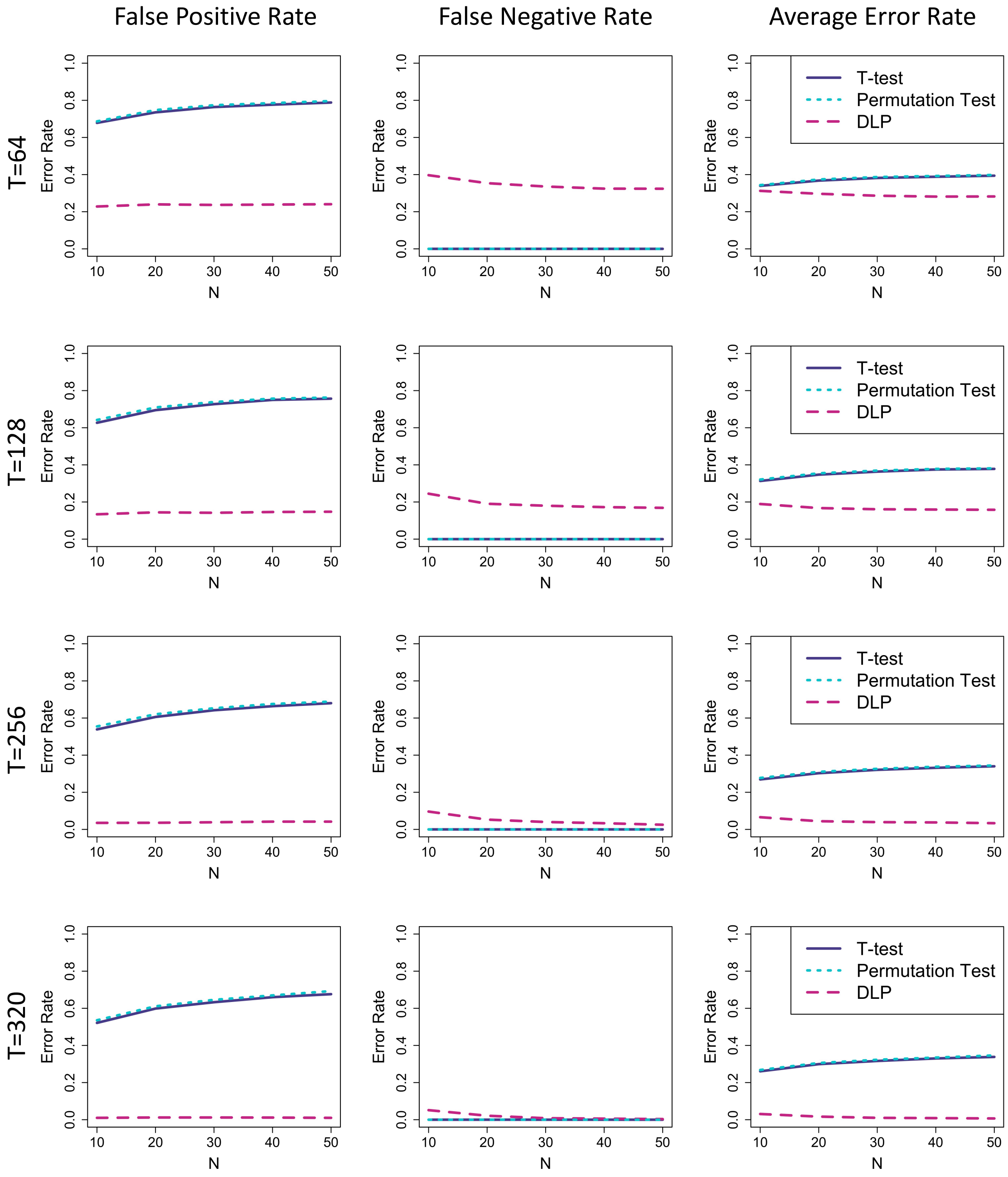

Figure 5.4: Results of the simulation study using the DLP. Rows represent different time series lengths and columns represent the different errors. The first column shows the error rate among truly null ROIs (false positive). The second column shows the error rate among truly connected ROIs (false negative). The third column is the average of the first two columns. The DLP approach maintains a much smaller false positive rate than the traditional methods (which show very similar behavior across all simulations. For T=64 and T=128, we see an increased false negative rate for the DLP approach, where the other two have extremely small errors. The DLP results are extremely similar to the LP approach (Figure 1) due to the rapid shrinking of the inconclusive region as the statistical information increases.

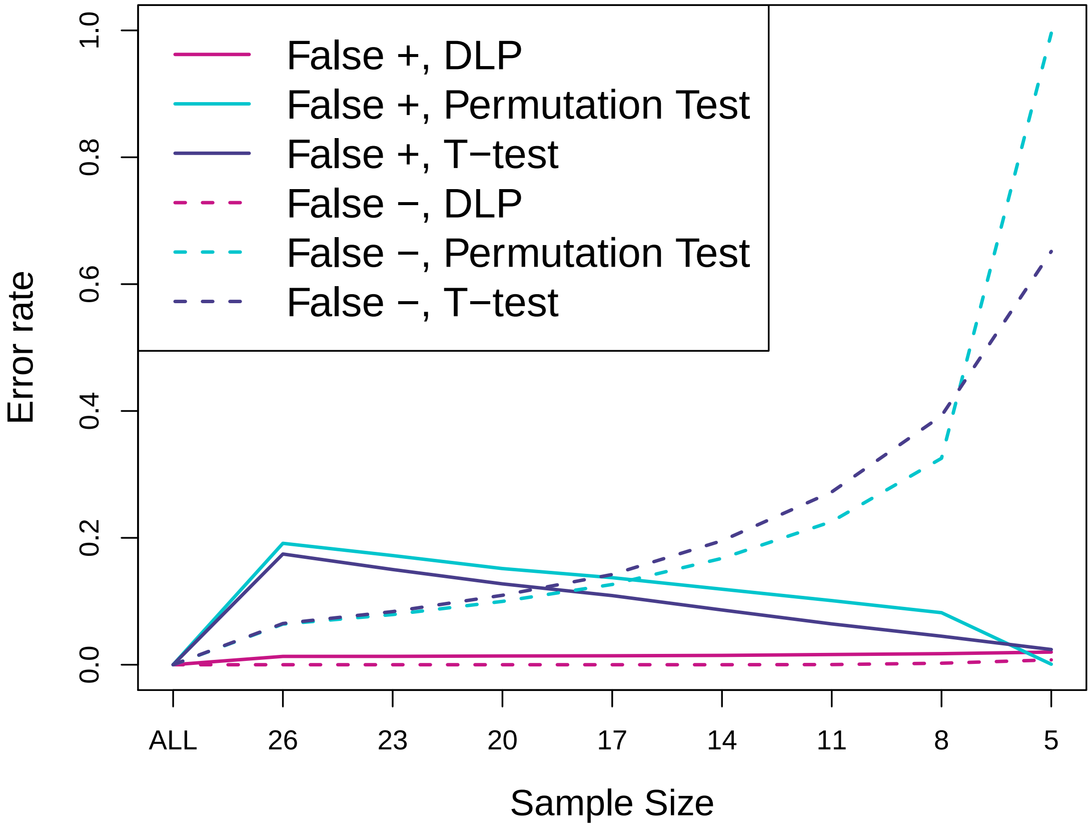

Figure 5.5: Data decimation results. The likelihood paradigm approach maintains very low Type I and Type II error rates down to a sample size of 8 subjects. The t-test and permutation test results, however, show higher Type I and Type II error rates across all sample sizes. As the sample size decreases, the frequentist approaches become less likely to reject a truly null pair of ROIs and more likely to fail to reject a truly connected pair.

CHAPTER 6

CONCLUSION

6.1 Summary

This dissertation aims to narrow the gap between the modern statistical analysis and medical image analysis. We focus on methods that improve inference and allow for researchers to get the most out of their data. In Chapter 2 we proposed the use of simulation extrapolation (SIMEX) and the wild bootstrap to estimate the bias and variance, respectively, of generalized fractional anisotropy (GFA) in high angular resolution diffusion imaging (HARDI) data. Chapter 3 provided an application of the methodology detailed in Chapter 2 to a study in which traveling subjects were imaged multiple times on multiple independent scanners. We calculated the bias and variance of GFA for each scan and used these values to learn about the quality of our data. We also provided an example workflow to instruct researchers on the use of these methods for measuring data quality and choosing an appropriate model for inference. While the methods proposed in Chapter 2 work well, they tend to be computationally intensive and time consuming. In order to speed up this process, we introduced a deep learning approach to bias and variance estimation in Chapter 4. In Chapter 5 we changed gears and switched from DW-MRI to fMRI. We introduced novel methodology for the identification of functionally connected regions of interest via an application of the likelihood paradigm to resting-state fMRI data. This technique is shown to outperform traditional frequentist techniques in terms of average error rates.

These contributions focused on improving inference through understanding of the statistical properties of medical image data. Several chapters address the statistical properties of DW-MRI metrics and provide two distinct methods for estimating such properties. The final chapter details an inferential technique for analysis of resting-state fMRI data.

6.2 Bias and Variance Estimation on HARDI

Section I (Chapters 2-4) of this dissertation focused on the estimation of the bias and variance of generalized fractional anisotropy (GFA) in high angular resolution diffusion imaging (HARDI). First, we presented a method for the estimation of bias through an extension of the simulation extrapolation (SIMEX) technique and variance via a wild bootstrap technique. These methods proved to be quite useful in estimating these values for single acquisitions of HARDI data. Before this dissertation, the estimation of bias and variance had only previously been shown for diffusion tensor imaging (DTI) (Lauzon et al. (2013)). These bias and variance estimates can be used for data quality assurance as shown in Chapter 3. Chapter 3 provided a real-world application of these techniques to a study involving multiple repeated image acquisitions across several subjects. The results of the study demonstrate the importance of a review of scan quality before combining data from several studies, thus ensuring that the bias and variance are compatible across the entire study. The statistical techniques provided in Chapter 2 rely on Monte Carlo simulations, which work well for small samples of scans, but do not generalize well to large studies or pipeline inclusion due to the time required for each dataset. Thus, we created deep learning models that can effectively estimate these bias and variance values without requiring lengthy simulations (Chapter 4). These deep networks are perfect for inclusion in quality assurance pipelines as a way to determine scan quality both quantitatively and quickly.

Together, these methods can be used for the quantitative comparison of scanners, processing techniques, and analysis methods that would not have been possible previously. This would allow for better informed choices of methodology for future studies, resulting in better research and increased innovation.

6.3 rs-fMRI Data Analysis via the Likelihood Paradigm

Section II (Chapter 5) of this dissertation shifted the focus from DW-MRI to resting-state functional MRI. In this chapter we introduced a technique for the identification of functionally connected areas of the brain via an application of the likelihood paradigm to rs-fMRI data. The proposed technique allows for the control of both Type I and Type II error, resulting in improved inference when compared to traditional frequentist techniques.

6.4 Summary of Contributions

The final contributions of this dissertation to the fields of statistics and medical image processing are summarized below.

- We extended the use of SIMEX and the wild bootstrap for bias and variance estimation of GFA for HARDI data. These techniques can be used for bias-correction, or for quantitative comparisons across subjects, sites, or scanners.

- We provided a sample workflow for the evaluation of inter-site bias and variance of GFA in HARDI. This workflow can be used to understand the statistical properties of the acquired data and to inform model selection for research contrasts. The workflow is easy to use and provides clear steps that any researcher can follow.

- We developed a collection of deep learning networks to allow for estimation of bias and variance of GFA for HARDI data approximately 200x faster than the statistical techniques introduced in Chapter 2. The speed of these techniques allows for their inclusion in pipelines, which will allow for their use in real-time scan environments.

- We provided methodology for the detection of functionally connected brain areas using the likelihood paradigm applied to rs-fMRI data. The use of our methodology allows for lower average error rates in identifying these areas as demonstrated on simulated data and clinical data.

REFERENCES

Alexander, A. L., Lee, J. E., Lazar, M. and Field, A. S. (2007), Diffusion tensor imaging of the brain, *Neurotherapeutics* **4**(3), 316–329.

Anderson, A. W. (2001), Theoretical analysis of the effects of noise on diffusion tensor imaging, *Magnetic Resonance in Medicine* **46**, 1174–1188.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S. and Hyde, J. S. (1992), Time course epi of human brain function during task activation, *Magnetic Resonance in Medicine* **25**(2), 390–397.

Basser, P. J. (1997), Quantifying errors in fiber-tract direction and diffusion tensor field maps resulting from MR noise, *Proceedings of the 5th Annual Meeting of ISMRM, Vancouver, Canada 1997* .

Basser, P. J. and Jones, D. K. (2002), Diffusion-tensor mri: theory, experimental design and data anlysis - a technical review, *NMR Biomed* **15**, 456–467.

Basser, P. J., Mattiello, J. and LeBihan, D. (1994*a*), Estimation of the effective self-diffusion tensor from the NMR spin echo, *Journal of Magnetic Resonance, Series B* **103**(3), 247–254.

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

Basser, P. J. and Pajevic, S. (2000), Statistical artifacts in diffusion tensor MRI (DTMRI) caused by background noise, *Magnetic Resonance in Medicine* **44**, 41–50.

Basser, P. J. and Pierpaoli, C. (1996), Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri, *Journal of Magnetic Resonance, Series B* **111**, 209–219.

Bastin, M. E., Armitage, P. A. and Marshall, I. (1998), A theoretical study of the effect of experimental noise on the measurement of anisotropy in diffusion imaging, *Magnetic Resonance Imaging* **16**, 773–785.

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 Series B Statistical Methodology* **57**(1), 289–300.

Biswal, B., Yetkin, F. Z., Haughton, V. M. and Hyde, J. S. (1995), Functional connectivity in the motor cortex of resting brain using echo-planar mri, *Magnetic Resonance in Medicine* **34**(4), 537–541.

Blume, J. D. (2002), Likelihood methods for measuring statistical evidence., *Statistics in Medicine* **21**, 2563–2599.

Carroll, R. J., Kuchenhoff, F., Lombard, F. and Stefanski, L. A. (1996), Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models, *Journal of the American Statistical Association* **91**(433), 242–250.

Chang, L. C., Koay, C. G., Pierpaoli, C. and Basser, P. J. (2007), Variance of estimated DTI-derived parameters via first-order perturbation methods, *Magnetic Resonance in Medicine* **57**, 141–149.

Chen, J. E. and Glover, G. H. (2016), Functional magnetic resonance imaging methods, *Neuropsychology Review* **25**(3), 289–313.

Chollet, F. et al. (2015), 'Keras', https://github.com/fchollet/keras.

Cook, J. R. and Stefanski, L. A. (1994), Simulation-Extrapolation Estimation in Parametric Measurement Error Models, *Journal of the American Statistical Association* **89**(428), 1314–1328.

Cook, P. A., Bai, Y., Nedjati-Gilani, S., Seunarine, K. K., Hall, M. G., Parker, G. J. and Alexander, A. C. (2006), Camino: Open-source diffusion-mri reconstruction and processing, *14th Scientific Meeting of the International Society for Magnetic Resonance in Medicine* 2759.

Cordes, D., Haughton, V., Carew, J. D., Arfanakis, K. and Maravilla, K. (2002), Hierarchical clustering to measure connectivity in fMRI resting-state data., *Magnetic Resonance Imaging* **20**(4), 305–317.

Descoteaux, M., Angelino, E., Fitzgibbons, S. and Deriche, R. (2007), Regularized, fast, and robust analytical Q-ball imaging, *Magnetic Resonance in Medicine* **58**(3), 497–510.

Di Martino, A., Yan, C. G., Li, Q., Denio, E. and Castellanos, F. X. e. a. (2014), The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, *Molecular Psychiatry* **19**(6), 659–667.

Efron, B. (1992), Bootstrap methods: another look at the jackknife, *in* S. Kotz and N. L. Johnson, eds, 'Breakthroughs in Statistics: Methodology and Distribution', Springer New York, New York, NY, 569–593.

Farrell, J. A., Landman, B. A., Jones, C. K., Smith, S. A., Prince, J. L., van Zijl, P. C. and Mori, S. (2007), Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 T, *Magnetic Resonance Imaging* **26**(3), 756–767.

Fonov, V. S., Evans, A. C., Almli, C. R. and Collins, D. L. (2009), Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, *NeuroImage* **47**.

Fonov, V. S., Evans, A. C., Botterton, K., Almli, C. R., McKinstry, R. C. and Collins, D. L. (2011), Unbiased average age-appropriate atlases for pediatric studies, *NeuroImage* **54**(1).

Gudbjartsson, H. and Patz, S. (1995), The Rician distribution of noisy MRI data, *Magnetic Resonance in Medicine* **34**(6), 910–914.

Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press.

Hainline, A. E., Nath, V., Parvathaneni, P., Blaber, J. A., Rogers, B., Newton, A., Luci, J., Edmonson, H., Kang, H. and Landman, B. A. (2018), Evaluation of intersite bias and variance in diffusion-weighted mri, *Proc.SPIE* **10574**.
URL: *https://doi.org/10.1117/12.2293735*

Hainline, A. E., Nath, V., Parvathaneni, P., Blaber, J. A., Schilling, K. G., Anderson, A. W., Kang, H. and Landman, B. A. (2018), Empirical single sample quantification of bias and variance in q-ball, *Magnetic Resonance in Medicine* **80**, 1666–1675.

Hainline, A. E., Nath, V., Parvathaneni, P., Schilling, K. G., Blaber, J. A., Anderson, A. W., Kang, H. and Landman, B. A. (2018), A deep learning approach to estimation of bias and variance in high angular resolution diffusion imaging. (in press).

Hainline, A. and Kang, H. (2018), Evidence-based inference on resting state functional connectivity. (in press).

Hayasaka, S. and Nichols, T. E. (2004), Combining voxel intensity and cluster extent with permutation test framework., *NeuroImage* **23**, 54–63.

Hess, C. P., Mukherjee, P., Han, E. T., Xu, D. and Vigneron, D. B. (2006), Q-ball reconstruction of multimodal fiber orientations using the spherical harmonic basis, *Magnetic Resonance in Medicine* **56**(1), 104–117.

Holmes, A. P., Blair, R. C., Watson, J. D. and Ford, I. (1996), Nonparametric analysis of statistic images from functional mapping experiments, *Journal of Cerebral Blood Flow Metabolism* **16**(1), 7–22.

Hutchinson, E. B., Avram, A. V., Irfanoglu, M. O., Koay, C. G., Barnett, A. S., Komlosh, M. E., Ozarslan, E., Schwerin, S. C., Juliano, S. L. and Pierpaoli, C. (2017), Analysis of the effects of noise, DWI sampling, and value of assumed parameters in diffusion MRI models, *Magnetic Resonance in Medicine* **78**(5), 1767–1780.

Jack, C. R., Bernstein, M. A., Fox, N C Thompson, P., Alexander, G., Harvey, D. and Borowski, B. e. a. (2014), The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, *Journal of Magnetic Resonance Imaging* **27**(4), 685–691.

Jones, D. K. (2008), Tractography gone wild: probabilistic fibre tracking using the wild bootstrap with diffusion tensor MRI, *IEEE Transactions on Medical Imaging* **27**(9), 1268–1274.

Kang, H., Blume, J. D., Ombao, H. and Badre, D. (2015), Simultaneous Control of Error Rates in fMRI Data Analysis., *NeuroImage* **123**, 102–113.

Kingma, D. P. and Ba, J. (2017), 'Adam: A method for stochastic optimization'.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S. and Turner, R. (1992), Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation, *Proceedings of the National Academy of Sciences of the United States of America* **89**(12), 5675–5679.

Lauzon, C. B., Asman, A. J., Crainiceanu, C., Caffo, B. C. and Landman, B. A. (2011), Assessment of bias for MRI diffusion tensor imaging using SIMEX, *Medical Image Computing and Computer Assisted Intervention* **14**(2), 107–115.

Lauzon, C. B., Crainiceanu, C., Caffo, B. C. and Landman, B. A. (2013), Assessment of bias in experimentally measured diffusion tensor imaging parameters using SIMEX, *Magnetic Resonance in Medicine* **69**(3), 891–902.

Lee, M. H., Smyser, C. D. and Shimony, J. S. (2013), Resting-state fmri: A review of methods and clinical applications, *American Journal of Neuroradiology* **34**(10), 1866–1872.

Li, S. J., Wu, G., Zhang, M. J., Franczak, M. and Antuono, P. G. (2002), Alzheimer disease: evaluation of a functional mr imaging index as a marker, *Radiology* **225**, 253–259.

Liu, R. Y. (1988), Bootstrap Procedures under some Non-I.I.D. Models, *The Annals of Statistics* **16**(4), 1696–1708.

Logothetis, N. K. and Wandell, B. A. (2004), Interpreting the bold signal, *Annual Review of Physiology* **66**, 735–769.

Lowe, M. J., Beall, E. B., Sakaie, K. E., Koenig, K. A., Stone, L., Marrie, R. A. and Phillips, M. D. (2008), Resting state sensorimotor functional connectivity in multiple sclerosis inversely correlates with transcallosal motor pathway transverse diffusivity, *Human Brain Mapping* **29**, 818–827.

Lowe, M. J., Phillips, M. D., Lurito, J. T., Mattson, D., Dzemidzic, M. and Mathews, V. P. (2002), Multiple sclerosis: low-frequency temporal blood oxygen level-dependent fluctuations indicate reduced functional connectivity initial results, *Radiology* **224**, 184–192.

Maier-Hein, K. H., Neher, P. F., Houde, J.-C., Cote, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y. C., Ji, Q., Reddick, W. E., Glass, J. O., Chen, D. Q., Feng, Y., Gao, C., Wu, Y., Ma, J., Renjie, H., Li, Q., Westin, C.-F., Deslauriers-Gauthier, S., Gonzalez, J. O. O., Paquette, M., St-Jean, S., Girard, G., Rheault, F., Sidhu, J., Tax, C. M. W., Guo, F., Mesri, H. Y., David, S., Froeling, M., Heemskerk, A. M., Leemans, A., Bore, A., Pinsard, B., Bedetti, C., Desrosiers, M., Brambati, S., Doyon, J., Sarica, A., Vasta, R., Cerasa, A., Quattrone, A., Yeatman, J., Khan, A. R., Hodges, W., Alexander, S., Romascano, D., Barakovic, M., Auria, A., Esteban, O., Lemkaddem, A., Thiran, J.-P., Cetingul, H. E., Odry, B. L., Mailhe, B., Nadar, M., Pizzagalli, F., Prasad, G., Villalon-Reina, J., Galvis, J., Thompson, P., Requejo, F., Laguna, P., Lacerda, L., Barrett, R., Dell'Acqua, F., Catani, M., Petit, L., Caruyer, E., Daducci, A., Dyrby, T., Holland-Letz, T., Hilgetag, C., Stieltjes, B. and Descoteaux, M. (2017), The challenge of mapping the human connectome based on diffusion tractography, *Nature Communications* **8**(1349).

Mason, D. R. and Schuenemeyer, J. H. (1983), A Modified Kolmogorov-Smirnov Test Sensitive to Tail Alternatives, *Annals of Statistics* **11**(3), 933–946.

MATLAB (2016), *version 9.0.0 (R2016a)*, The MathWorks Inc., Natick, Massachusetts.

Nath, V., Schilling, K. G., Blaber, J., Ding, Z., Anderson, A. and Landman, B. A. (2017), Comparison of multi-fiber reproducibility of pas-mri and q-ball with empirical multiple b-value hardi, *Proc SPIE Int Soc Opt Eng* **Proceedings Volume 10133**.

Nath, V., Schilling, K. G., Parvathaneni, P., Blaber, J., Hainline, A. E., Ding, Z., Anderson, A. and Landman, B. A. (2018), Empirical estimation of intravoxel structure with persistent angular structure and q-ball models of diffusion weighted mri, *Journal of Medical Imaging* **5**(1).

Nichols, T. E. and Hayasaka, S. (2003), Controlling the familywise error rate in functional neuroimaging: a comparative review., *Statistical Methods in Medical Research* **12**(5), 419–446.

Nichols, T. E. and Holmes, A. P. (2001), Nonparametric permutation tests for functional neuroimaging: A primer with examples, *Human Brain Mapping* **15**, 1–25.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H. and Ugurbil, K. (1992), Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging, *Proceedings of the National Academy of Sciences of the United States of America* **89**(13), 5951–5955.

Python Core Team (2015), *Python Language Reference*, Python Software Foundation. **URL:** *http://www.python.org/*

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. and Shulman, G. L. (2001), A default mode of brain function., *PNAS* **98**(2), 676–682.

Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, Chapman and Hall.

Skare, S., Li, T. Q., Nordell, B. and Ingvar, M. (2000), Noise considerations in the determination of diffusion tensor anisotropy, *Magnetic Resonance Imaging* **18**, 659–669.

Smith, S. M., Jenkinson, M., Woolrich, M. W. and Beckmann, C. F. e. a. (2004), Advances in functional and structural MR image analysis and implementation as FSL., *NeuroImage* **23**(S1), 208–219.

Stejskal, E. O. and Tanner, J. E. (1965), Spin diffusion measurements: Spin echoes in the presence of a field gradient, *The Journal of Chemical Physics* **42**, 288–292.

Tuch, D. S. (2004), Q-ball imaging, *Magnetic Resonance in Medicine* **52**(6), 1358–1372.

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**(4), 577–582.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. and Joliot, M. (2002), Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain, *NeuroImage* **15**(1), 273–289.

Wang, K., Jiang, T., Liang, M., Wang, L., Tian, L., Zhang, X., Li, K. and Liu, Z. (2006), Discriminative analysis of early alzheimer's disease based on two intrinsically anti-correlated networks with resting-state fmri, *Med. Image Comput. Comput. Assist. Interv. Int. Conf. Med. Image Comput. Comput. Assist. Interv.* **9**, 340–347.

Weissenbacher, A., Kasess, C., Gerstl, F., Lanzenberger, R., Moser, E. and Windischberger, C. (2009), Correlations and anticorrelations in resting-state functional connectivity mri: A quantitative comparison of preprocessing strategies, *NeuroImage* **47**(4), 1408–1416.

Whitcher, B., Tuch, D. S., Wisco, J. J., Sorensen, A. G. and Wang, L. (2008), Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging, *Human Brain Mapping* **29**(3), 346–362.

Zhou, Y., Shu, N., Liu, Y., Song, M., Hao, Y., Liu, H., Yu, G., Liu, Z. and Jiang, T. (2008), Altered resting-state functional connectivity and anatomical connectivity of hippocampus in schizophrenia, *Schizophrenia Research* **100**, 120–132.