JYU DISSERTATIONS 93

Jia Liu

Data Augmentation under Rician Noise Model in Diffusion MRI with Applications to Human Brain Studies

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF MATHEMATICS
AND SCIENCE

JYU DISSERTATIONS 93

Jia Liu

Data Augmentation under Rician Noise Model in Diffusion MRI with Applications to Human Brain Studies

Esitetään Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen salissa Alfa
kesäkuun 5. päivänä 2019 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Mathematics and Science of the University of Jyväskylä,
in building Agora, hall Alfa, on June 5, 2019 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2019

Editors:

Salme Kärkkäinen
Department of Mathematics and Statistics
University of Jyväskylä
Finland

Timo Hautala
Open Science Centre
University of Jyväskylä
Finland

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7787-0

ISBN 978-951-39-7787-0 (PDF)
URN:ISBN:978-951-39-7787-0
ISSN 2489-9003 Jyväskylä University Printing House, Jyväskylä 2019

ABSTRACT

Liu, Jia
Data Augmentation under Rician Noise Model in Diffusion MRI with Applications to Human Brain Studies
Jyväskylä: University of Jyväskylä, 2019, 83 p.
(JYU Dissertations
ISSN 2489-9003; 93)
ISBN 978-951-39-7787-0 (PDF)

Diffusion magnetic resonance imaging (diffusion MRI) is capable of measuring the displacement diffusion of water molecules and of providing a unique insight by means of image contrasts from measurements to probe non-invasively the microscopic anatomical architectures of organic tissues in vivo. Many diffusion imaging approaches have been developed to measure the underlying diffusion function, among which diffusion tensor imaging (DTI) is the most popular. A conventional modeling approach postulates a Gaussian displacement distribution at each voxel characterized by means of a second order symmetric and positive definite diffusion tensor. Typically, the inference is based on a linearized log-normal regression model. However, such an approximation fails to fit the high frequency and/or the low signal to ratio (SNR) measurements containing important information on water diffusion. The diffusion weighted MR measurements are sparse and noisy, and after the Fourier inversion they yield a non-linear regression problem. However, working with the non-linear model for the data directly leads to heavy computation.

In this thesis, I present a series of novel statistical methodologies to solve the computational problem. By using data augmentation, the non-linear regression problem under the Rician noise model is reduced to the generalized linear modeling (GLM) framework. For different purposes, we use both Bayesian and frequentist statistical inferences: A Bayesian hierarchical model is established to estimate the marginal posterior distribution of every parameter of interest, where we apply the Markov chain Monte Carlo (McMC) method, exploring the state space to compute averages under the joint posterior distribution of the unknown parameters and latent variables. Moreover, we also implement Variational Bayes (VB) algorithms as a faster scheme for converging to the optimum of each posterior distribution. Under the Bayesian framework, a regularization technique is developed for modeling the contextual dependence (interaction) between the tensors. This is done by constructing an isotropic prior for the tensor fields through the Gaussian Markov random fields (GMRF). This model is intended to smooth and denoise the image. In terms of computational issues in practice, we further employ the expectation-maximization (EM) algorithm under the joint likelihood in GLM by the data augmentation to reduce computational burden in both Bayesian and frequentist frameworks. This deterministic algorithm is implemented under the assumption of voxel independence in both maximum a posterior (MAP) estimation and maximum likelihood estimation (MLE). Furthermore, we apply the stabilized Fisher scoring method for achieving fast convergence in the calculation of the tensor parameter. In addition, we address the essential difference between these two inferences working in dMRI. All these methodologies are described in four papers under several popular signal decay models in dMRI, implemented and experimented both with synthetic and real data of the human brain, and compared with different popular methods in dMRI and in the recent literature.

Keywords: Bayesian regularization, diffusion magnetic resonance imaging, diffusion tensor imaging, constrained diffusion kurtosis imaging, expectation maximization, Fisher scoring, generalized linear modeling, high angular resolution diffusion imaging, positivity condition, Markov chain Monte Carlo, maximum likelihood estimation, Rician noise, spherical harmonics, statistical inference, variational Bayes

Author Jia Liu

Department of Mathematics and Statistics
University of Jyväskylä
Finland

Supervisors Docent Dario Gasbarra

Department of Mathematics and Statistics
University of Helsinki
Finland

Doctor Salme Kärkkäinen

Department of Mathematics and Statistics
University of Jyväskylä
Finland

Professor emeritus Antti Penttinen

Department of Mathematics and Statistics
University of Jyväskylä
Finland

Reviewers Associate Professor Santiago Aja-Fernández

ETSI Telecomunicación- Laboratorio de Procesado de imagen, Universidad de Valladolid
Spain

Professor emeritus Lasse Holmström

Research Unit of Mathematical Sciences
University of Oulu
Finland

Opponent Professor Ian Dryden

School of Mathematical Sciences
University of Nottingham
United Kingdom

PREFACE

In this thesis we present new developments of statistical methodologies in both frequentist and Bayesian inference framework for estimating tensor-derived quantities under different dMRI protocols, and to illustrate significantly promising results through relevant comparison on both synthetic and real data to study the human brain. The aim of the work is to solve the problems in dMRI and to contribute the exploration of the structure of the human brain which may impact on the diagnosis of brain diseases in clinical applications.

My doctoral work is composed of four papers describing different novel statistical methods and developments implemented in dMRI with an object of the human brain.

Terminology used in the summary

dMRI – diffusion magnetic resonance imaging

DTI – diffusion tensor imaging

DKI – diffusion kurtosis imaging

dMRI – diffusion magnetic resonance imaging

DW-MRI – diffusion weighted magnetic resonance imaging

EM – expectation-maximization

GLM – generalized linear modeling

GMRF – Gaussian Markov random fields

HARDI – high angular resolution diffusion imaging

MAP – maximum a posterior

McMC – Markov chain Monte Carlo

MLE – maximum likelihood estimation

SNR – signal to noise ratio

VB – variational Bayes

WLS – weighted least squares

Author's contributions This thesis consists of a summary and four original research papers. The author's contributions of each paper are listed below.

PI The author jointly worked in method development and implementation, and in drafting the manuscript. The author designed and implemented the simulation study, and interpreted the results in the paper.

PII The author developed and implemented the methods, drafted the manuscript, interpreted the results and critically revised the manuscript.

PIII and **PIV** The author developed and implemented the method, drafted the manuscripts and interpreted the results. The author critically revised **PIII** after the peer-review.

ACKNOWLEDGEMENTS

I would like to express my great gratitude to my supervisor Dario Gasbarra, who introduced this research topic to me and has shared a lot of his time with me to carry out this project. I also thank Juha Railavo for his precious collaboration; he introduced DTI to Dario and me and organized the data collection.

I am deeply indebted to Emeritus Professor Antti Penttinen for his huge encouragement and inspiring guidance, especially at the last stage of my PhD studies. He has invested a lot of his time in this project, been relentless in reading and made many substantive suggestions and insightful comments that completely improved the corrections, clarity and readability of this work. Without his counsel and follow-up, I would not finish this work smoothly nor be aware of the importance of independence in doing research. I would also like to thank my co-supervisor Doctor Salme Kärkkäinen. She has given me a lot of help during my PhD in Jyväskylä and has never mind sharing her experience at different stages of my studies. I would be grateful for the two pre-examiners for their expert comments.

My thanks also spread to the Department of Mathematics and Statistics, University of Jyväskylä and to every colleague there for the nice research environment and kind discussions on research and other aspects. As an international student and member of staff, I would like to specially thank Tuula Blåfield for her hard work in English checking of the summary and also Sari Eronen, Hannele Säntti-Ahomäki and Eeva Partanen, for their help in different practical things.

Further, I would like to thank my previous advisor Assistant Professor Jarno Vanhatalo from University of Helsinki for his kind support. My thanks also to Doctor Viljami Sairanen from the Department of Physics, University of Helsinki, who commented on Chapter 2 of my thesis and Doctor Daniel Blande from University of Eastern Finland who helped me correct the English of several chapters of this thesis.

I wish to thank the CSC-IT Center for Science Ltd. for providing powerful computing resources. This work was funded by the Doctoral Program in Computing and Mathematical Sciences (COMAS) and the Department of Mathematics and Statistics, University of Jyväskylä.

Finally, I am extremely grateful to my family and my parents for their love and constant support throughout my life. This work is dedicated to my daughter, Kangxin Päivi.

Jia Liu

May 2019, Helsinki, Finland

CONTENTS

LIST OF FIGURES

LIST OF INCLUDED ARTICLES

PI Dario Gasbarra, Jia Liu and Juha Railavo. Data augmentation in Rician noise model and Bayesian Diffusion Tensor Imaging. *Submitted*, (2019).

PII Jia Liu, Dario Gasbarra and Juha Railavo. Fast estimation of diffusion tensors under Rician noise by the EM algorithm. *Journal of Neuroscience Methods, 257: 147-158*, (2016).

PIII Jia Liu. An improved EM algorithm for solving MLE in constrained diffusion kurtosis imaging of human brain. *Submitted*, (2019).

PIV Jia Liu, Dario Gasbarra and Juha Railavo. Variational Bayes Estimation in Constrained Kurtosis Diffusion Imaging under a Rician Noise Model. *Submitted*, (2019).

PART I

1 INTRODUCTION

1.1 Background

Human brain is one of the most interesting and mysterious media in the world. The living brain is soft and delicate and is protected by the solid bones of the skull and covered by a thick layer of neural tissue termed cerebral cortex. An adult's brain takes about 2% of the body weight being around 1.5kg on an average. At the macroscopic level, the brain is divided into grey matter (GM) and white matter (WM): 40% of the brain is occupied by GM, and the remaining is filled by WM. The human brain is composed of 73% water, of which WM structures (see Figure 1a) have significantly higher myelin water percentages than the GM structures, and the water diffuses preferentially along the fibre bundles, see Goss (1960), Damasio (1995), Mai et al. (1997), Thomalla et al. (2005) for deeper knowledge of the anatomy of human brain. Therefore, in order to explore the brain architecture represented by the fine and rich fibrous structure (see Figure 1b), it is extremely important to study *in vivo* water diffusion, especially how it maps the white matter pathways.

Diffusion magnetic resonance imaging (diffusion MRI) is a currently known imaging method that enables us to measure the diffusion of water molecules and to probe the microstructure of the brain by the measurements (Tuch et al. , 2003). It provides a powerful non-invasive way to retrieve the anatomical and connectivity information of the brain. This information resource is thought to be useful and had/may have strong impacts on clinics for diagnosis of brain disorders, such as ischemic stroke (Mori , 2007, Thomalla et al. , 2005) and dementia with Lévy bodies (DLB), see Kantarci (2010), Mak et al. (2014).

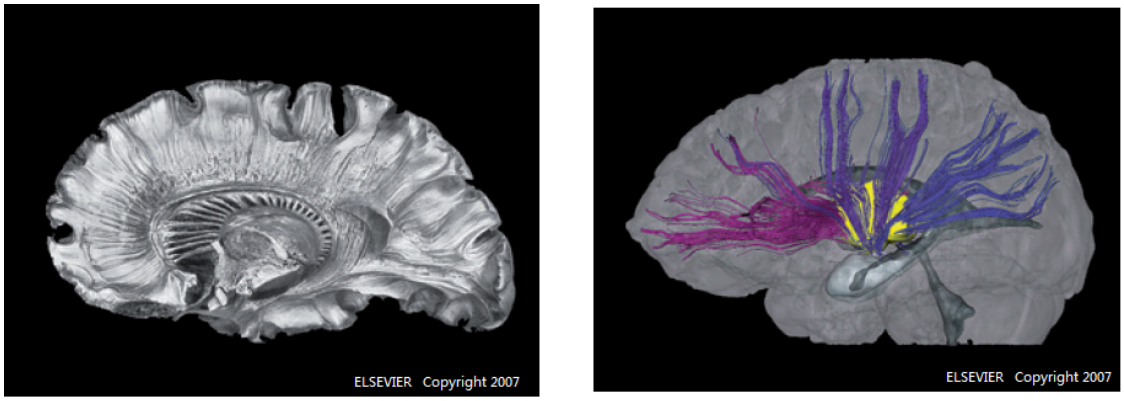


FIGURE 1 (a) The macroscopic WM architecture from a postmortem human sample. (b) DTI-based reconstruction. The figures are reprinted from Mori (2007). Copyright (2007), with permission from Elsevier, *http://www.elsevier.com.*

1.2 Motivation

The diffusion MRI measurements are sparse and noisy, especially when the b-value is large, see Figure 2. A common simplifying assumption about the dif-

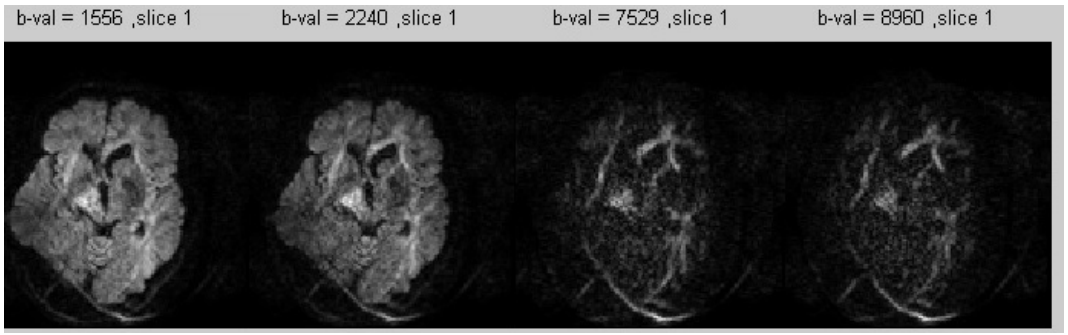
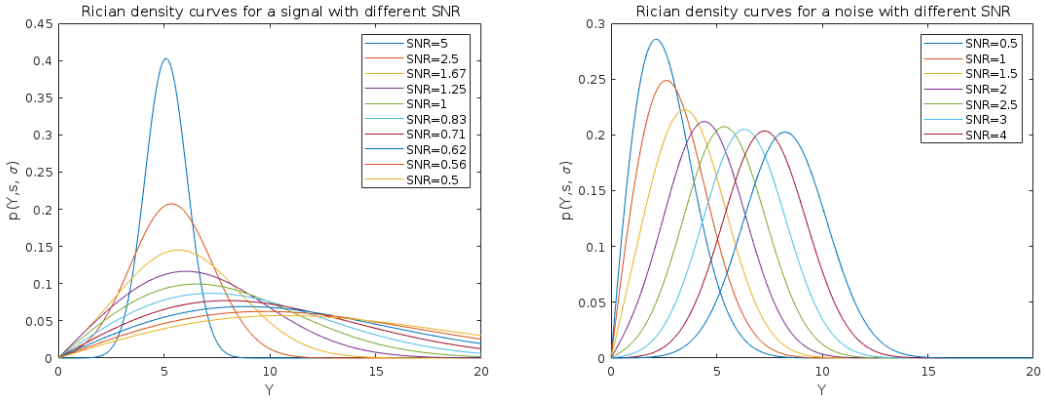


FIGURE 2 Diffusion weighted MR images generated by applying different b-values.

fusion magnetic resonance (MR) signal measurements (the data) is a Gaussian distribution. This assumption fits data well in those particular cases where the signal-to-noise ratio is (SNR $=S/\sigma$) $\geq$ 3. However, statistical inference reveals that the information retrieved in this way is unreliable in the low SNR regime. This is because the probability distribution of the data is far from being Gaussian, see Figure 3. In the past decades, numerous works such as Henkelman (1985), Bernstein et al. (1989), Andersen (1996), have been devoted to the study of the effects of the complex Gaussian noise in magnitude MRI. The theoretical distribution of the magnitude data has been proved to have a Rician distribution (Jones and Basser , 2004, Henkelman , 1985, Zhu et al. , 2007, Assemlal et al. , 2009, Landman et al. , 2007). Several authors, such as Zhu et al. (2007), Salvador et al. (2004), add the noise-induced bias into the measurements so that a simple



(a) (b)

FIGURE 3 The Rician density curves for a signal $S = 5$ (Figure 3a) and for standard variance σ from the noise (Figure 3b) of different SNR. For SNR ≥ 3, the density is approximated to be Gaussian. Figure 3a shows that when the noise increases, the SNR decreases from the top curve to the bottom. Figure 3b depicts that for a fixed value of $\sigma = 2$, the SNR increases when the signals increase, and the distribution gradually tends towards the Gaussian distribution.

Gaussian noise model can be fitted to the data. But none of these techniques applies satisfactorily in the high b-value range and/or in the low SNR regime, and the Gaussian model does not fit the corrected data, see e.g. Mori (2007), Burdette et al. (2001). Furthermore, the Rician noise model (a commonly used name, referring to the complex Gaussian noise model with magnitude measurements, see Cardenas-Blanco et al. (2007)) has been used in the literature, e.g. Gudbjartsson and Patz (1995), Veraart et al. (2011), Andersson (2008), Lauwers et al. (2010), but in all cases the methods dealing with the Rician model are computationally intensive. On the other hand, Lu et al. (2006a) demonstrated in their experiments that the non-Gaussian behavior turned out to be more evident using diffusion-weighted sequences with higher b-values. The definition of the b-value can be found in Equation (2.10).

Statistical modeling and efficient computing are needed in combination to truly benefit from the full power of the noisy measurements of the diffusion spectra. Our motivation is that even a slight improvement of the statistical analyses and/or mathematical modeling could have significant impacts on neuroscience and on the diagnosis of common neurological disease of the brain. On the other hand, Dementia with Lévy bodies (DLB) is a demanding application and another strong motivation of this thesis. It is a brain disorder related to Parkinson's disease and is difficult to be diagnosed. Lévy bodies are found not only in the deep grey matter structures, but also diffusely in the brain cortex. Lévy bodies are spherical intraneuronal protein aggregates that consist primarily of alfa-synuclein, a presynaptic microtubule protein, see e.g. Brown, D.F. (1999), Issidorides et al. (1991). This is why DLB is considered to be a group of synucleopathy, and a group of disorders with alfa-synuclein gene mutations. Nowadays

DLB is regarded as the second among the most common neurodegenerative diseases, and is also found in other brain disorders such Alzheimer's disease and Parkinson's disease dementia (Alzheimer's Association , 2015). In the US, approximately 1 million individuals of the population have been diagnosed with dementia, which typically begins at age over 50, see e.g. NIA and NINDS (2015). Standard MRI findings are generally nonspecific despite the often prominent visual symptoms that characterize DLB. There are some studies, e.g. Firbank et al. (2007), Watson et al. (2012), in which DTI shows increased mean diffusivity in the amygdala region and especially in the inferior longitudinal and in the inferior fronto-occipital fasciculus. This is among the main reasons to carry out this work for a possible contribution to clinic practice through characterizing the quantities of diffusion.

1.3 Outline of the thesis

This thesis is composed of an introduction and four papers describing different novel statistical methods and developments implemented in diffusion MRI with applications to a human brain study. The introduction is divided into three parts. Part 1 (Chapters 2-3) provides the requisite background, aiming to help the readers who are not familiar with the research topic of this thesis. In Chapter 2, we briefly review the principle of MRI and introduce the diffusion and diffusion MRI. The diffusion tensor imaging (DTI) as a key term applied in the thesis as well as its extensions are introduced in Chapter 3. In Chapter 4, we describe the Rician model of the diffusion MR signal measurements, and highlight the regression problem in DTI. We then go through the commonly applied statistical methods and optimization tools for inferring diffusion through diffusion tensor estimation. Part II consists of Chapters 5-6, which contain the original contributions of this thesis. In Chapter 5 we derive two new ideas of data augmentation working with Rician likelihood of the data. Based on that, we propose an expectation-maximization (EM) algorithm for tensor estimation. A Markov chain Monte Carlo (McMC) and a Variational Bayes (VB) methods are two other novel alternatives in Bayesian framework, which are introduced in the following chapter. Additionally, we present a new development of imaging regularization for imaging smoothness. Part III consists of a conclusion about the methods developed, and Chapter 7 includes the comparison of the proposed methods and a short introduction of each paper included.

2 PRINCIPLES OF DIFFUSION MRI

2.1 Basics of MRI

Magnetic resonance imaging (MRI) has been extensively used to study the anatomy and the disease process of the living body. This imaging method is a practical application for a physical phenomenon known as nuclear magnetic resonance (NMR), see Das (2015), in which radio frequency (RF) pulses are used to perturb nuclei spins from a known equilibrium induced by an external magnetic field. As the amount of energy that a nucleus can absorb depends on the strength of the external magnetic field, slight deviations induced with spatial gradient fields can be used to determine the locations in which nuclei are excited. Excited nuclei, i.e., those that absorbed the RF pulse, rapidly relaxate towards equilibrium after the RF pulse ends and emit the excess energy as a measurable RF signal. The MR image is finally formed by coupling these measurements with the known spatial encoding. Even though the exact details of NMR physics and MRI engineering are certainly interesting topics, they are omitted as they lie beyond the scope of this dissertation. The readers who are interested in these topics can find more detailed description in, for instance, Hornak (1996), Slichter (2013), Brown et al. (2004). In MRI, the typical used nucleus refers to the single proton from a hydrogen atom due to its richness in water. This implicitly states that the proton density (the water concentration) dominates the MRI signal intensity.

In an external magnetic field B, protons have a microscopic magnetization and their precession is analogous to tiny spinning tops wobbling. The resonance (formally named as the Larmor frequency) refers to the rate of the precession. The magnetic field gradient is a technology introducing gradients which orientate in three directions x (right-left), y (up-down) and z (front-back) and generate a magnetic field that varies spatially according to the locations of the spins. The (static) magnetic field is known the B_0 field defined along the direction of z axis. The B_0 strength is measured in unit teslas (T). The relationship between the frequency (ω) and B_0 is formulated by the famous Larmor Equation (see for instance

Hashemi et al. (2012)), that is

$$\omega = \gamma B_0, \tag{2.1}$$

where the constant γ is the gyromagnetic ratio and ω denotes the Larmor frequency. The strength of magnetic field can be modulated linearly along each axis by applying the field gradient $G_{X_t} = (\partial B/\partial x, \partial B/\partial y, \partial B/\partial z)$ combined from the $x-$, $y-$ and $z-$ axis gradients, affecting at the frequency when the protons process.

In an MRI experiment, the water molecules start to process at the same frequency as B_0 is kept as homogeneous as possible, shown in the graph demonstration in Figure 4a. When the field gradient is applied, the water molecules at different locations experience different external magnetic field B and give different frequencies (the sine waves) as depicted in Figure 4b and c, where the water molecule sees stronger B, resonating at a higher frequency in Figure 4b than the one in c. The arrows in the circles describe the magnetic moment of three water molecules (counted in column) in the rotated frame at different locations, that is, the *phases* of the MR signals from each molecule. The water molecules in all plots

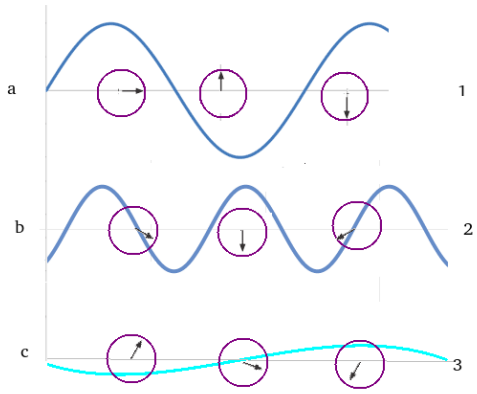


FIGURE 4 Three scenarios of three water molecules (counted in column) in the magnetic field without (Figure 4a) and with field gradient (Figure 4b and c). The locations of the particles determine their received values of B and the resonated frequencies vary. Because of the Larmor principle, different parts of the sample would have different resonance frequencies, and so a given resonance frequency could be associated with a given position.

of Figure 4 have a Larmor frequency of which Figure 4b and c are slightly different than that in Figure 4a described in Eq. 2.1. The Larmor frequency in Figure 4b and c has a general expression

$$\omega(x, y, z) = \gamma B_0 + \gamma G_{X_0} X_0, \tag{2.2}$$

where X_0 denotes the location of a water molecule. In MRI, the amount of the frequency difference is often considered as γB_0 is a constant. Therefore, in general the phase (ϕ) of the detected MR signal refers to the amount of phase difference,

which is given by

$$\triangle\phi(x) = \int \triangle\omega(t)dt = \gamma \int G(t)x(t)dt \qquad \text{and} \qquad \triangle\omega(t) = \gamma G(t)x(t), \tag{2.3}$$

see NessAiver and Moriel (1997), Haacke (1999), where $G(t)$ is the time-dependent magnetic field gradient and $x(t)$ denotes the trajectory of a particle at time t.

The spin-echo sequence Among the MRI pulse sequences, the spin-echo sequence is the most widely used one. It makes up of a series of events: a 90 ° RF pulse followed by one or more 180 ° refocusing pulses. The essential parameters in a spin echo sequence are the repetition time (RT) and the echo time (TE), where TR is the time interval between two consecutive 90 ° RF waves and TE is the time interval between the initial 90 ° RF pulse and the echo. TR depends on longitudinal relaxation time (T1) and TE depends on transverse relaxation time (T2), where T1 and T2 are known as basic parameters to determine the MR signal intensity (Hornak , 1996, Mori , 2007, Das , 2015). The T1 relaxation time measures how quickly the precession of the protons (the sum of the magnetic moment) recovers to its ground state in the direction of B_0. It depends on the B_0 field, in general: the higher B_0 field the longer period of T1. T2 relaxation refers to the progressive dephasing of spinning dipoles following the 90 ° pulse as seen in a spin-echo sequence. T1 and T2 relaxation rates affect signal to noise ratio (SNR) in an MR image, see Bloch (1946), Hesselink (1996). Relaxation refers to the process in which spins release the energy received from a RF, hence describes how signal changes with time, see NessAiver and Moriel (1997).

Three common spin-echo pulse sequences include the T1 weighted, the T2 weighted and their mixture. The T1 weighted sequences are obtained by using short TR and short TE values (typically TR $< 1000ms$, TE $< 30ms$) compared to the T2 sequences with long TR and long TE values, see Hornak (1996), NessAiver and Moriel (1997). A (T1 and T2) mixed sequence usually refers to the proton density weighted sequence, obtaining by a long TR and short TE sequence.

2.2 Diffusion

Diffusion usually refers to a process where water molecules or atoms in a liquid or a gas travel from a high concentration to a low one. This phenomenon is typically accompanied by a random thermal motion of the particles called Brownian motion (Mori , 2007, Tuch , 2002). When a water molecule travels in liquid water, the diffusion process (Itô diffusion) of the water molecule is described by the Itô stochastic differential equation (Baz and Chacko , 2004, Øksendal , 2003),

$$dX_t = a(X_t)d\xi_t, \tag{2.4}$$

where ξ_t is the 3-dimensional (3D) Brownian motion (BM) with zero drift and unit volatility. The random variable X_t describes the spatial position of the water

molecule determined at time t in 3D physical world and, correspondingly, $a(x)$ is a 3×3 matrix depending on the diffusing location of the water molecule, and $D_x = \frac{1}{2}a(x)a(x)^\top$ is usually called the apparent diffusion coefficient.

Free diffusion Multiple water molecules perform BM in a space determining the distribution of water molecules in free diffusion. When the space has no additional boundary conditions, water diffuses freely. Such kind of diffusion is called *free diffusion*. Owing to the central limit theorem, the displacement $(X_t - X_0)$ of the water molecules (also called the ensemble-average diffusion propagator (EAP), see Tuch (2002), Descoteaux (2010)) has a Gaussian distribution with zero mean and covariance $aa^\top = 2D$ (Chandrasekhar , 1943, Stieltjes et al. , 2013), which is given by

$$p(x,t) = (4\pi t)^{-3/2}|D|^{-1/2}\exp\big(-\frac{x^\top D^{-1}x}{4t}\big) \tag{2.5}$$

and in the case of free diffusion

$$p(x,t) = (4\pi tD)^{-3/2}\exp\big(-\frac{\|x\|^2}{4Dt}\big) \tag{2.6}$$

with scalar diffusion coefficient D.

Restricted diffusion In the brain, water diffusion is restricted due to additional surface boundary conditions such as cell membranes, boundaries and other complex compartments. These boundary conditions control the diffusion restrictions, which is a common scenario encountered in a biological tissue. In a short time interval $[0, t]$, it is unlikely for a water molecule to hit the boundary, hence the diffusion is almost free, and the probability distribution of the displacement is approximately Gaussian. After a certain time, water diffusion becomes restricted, hence the restricted diffusion coefficient D is time-dependent. The probability distribution of diffusion is then not a Gaussian function and can be extremely difficult to be formulated due to the complex of microstructure based on the very limited knowledge on the brain. Hence, the probability distribution of diffusion is determined by many additional parameters rather than a single diffusion coefficient.

2.3 How to measure the diffusion

A brief history Diffusion as a physical phenomenon has been an essential part of the history and development of diffusion MRI. Let us skip the era of nuclear magnetic resonance (NMR), diffusion MRI goes back to the nineteen fifties when Hahn (1950) observed the effect of diffusion to spin-echoes and pointed out that diffusion of the spins would reduce the amplitude of the observed signal over

an inhomogeneous magnetic field. This finding is considered as a keynote in understanding the diffusion MRI. Four years later, Carr and Purcell (1954) studied the effect of diffusion on free precession, and Torrey (1956) modified the Bloch equations to include a diffusion term with a spatially varying magnetic field. Stejskal and Tanner (1965), in their seminal paper, introduced the pulsed gradient spin echo sequence and showed the potential of diffusion related signal attenuation to probe the motion of molecules and to define the diffusion coefficient. Lauterbur (1973) published his groundbreaking paper entitled "Image formation by induced local interactions: Examples employing nuclear magnetic resonance". Owing to this work, in the year 2003 he shared the Nobel Prize together with Sir Peter Mansfield who proposed the echo-planar technique by studying the mathematical properties of the MR signal (Mansfield , 1977). In the experiment Lauterbur superimposed a gradient on the static uniform magnetic field. He also pointed out that it is possible to measure molecular diffusion from the decay of the MR signal. Based on that, diffusion weighted magnetic resonance imaging (DW-MRI) was introduced by Le Bihan et al. (1986) measuring the displacement of protons. In 1996, Stejskal and Tanner (Stejskal and Tanner , 1965) originated the famous pulse gradient spin echo experiment (PGSE), where they point out that diffusion occurring between two different diffusion gradient pulses can be reflected by the magnitude decay of the spin echoes. The question comes out: How to measure the diffusion through MRI?

Diffusion weighted (DW-) MR images The MR signal intensity is currently known information that is usually used in measuring diffusion. When applying a bipolar gradients, and in addition to the spatial field gradient: The MR signal intensity becomes very sensitive to the diffusion of protons when applying two consecutive and opposite (bipolar) gradients, which leads to imperfect rephasing and signal decay (loss) between the two gradients. As an illustration we use Figure 6 to describe what happens: The 90° RF generates a transversal magnetization. At time S1, the purple and blue protons (particles) start to see the same B and resonate at the same frequency. When the first (positive) gradient is applied, the two protons at different locations see different B and resonate at different frequencies (S2). At time S3, the system regains the homogeneous B, but the phases of the signal (the small arrows in the circles) are different than that at S1, which may cause the signal loss, equaling to a sum of signals from all the protons, which is the sum of all the arrows in the circles (Mori , 2007, Hornak , 1996). If there is no diffusion, the 2nd (negative but with the same strength) gradient helps the protons to regain the same phase (the positions of the arrows) as in S1 at time S4. If there is diffusion present, the protons perform random walk and change their spatial locations at time S3. An illustration describing such a phenomenon is depicted in Figure 6 below the green line: The blue water molecule performs a random walk in S3 until it moves to the location shown in S4. It then causes the protons imperfectly rephased in S4. Therefore, S2 and S4 are called the dephasing and the rephasing stage, respectively.

In the dephasing stage, the locations of all the protons are recorded using

their signal phase. If diffusion happens in some protons, their phase may be disrupted resulting in imperfect rephasing in S4, which is different than others (no diffusion). It turns out that diffusion can be inferred by signal decay (S/S_0) detected from the imperfect rephasing among all the protons, and signal decay can be measured by the signal with (S) and without diffusion weighting (S_0). Here the diffusion weight refers to the values gradients that determined by the strength (G), the duration (δ) of the gradient and the duration ($\triangle$) between the two gradients. The applied gradients that cause the diffusion weighting are named diffusion-weighting gradients. The stronger and longer the gradients are, the farther the water molecules move, which then causes the stronger diffusion and much signal is lost. Figure 5 illustrates four DW-MR images by applying different gradients. It shows that when the gradients become more and more stronger and longer as indicated by the b-values, the images turn much noisier representing that the signal loss increases. There are many ways to achieve and manipulate the diffusion weighting for acquiring imaging data, we illustrate some commonly used methods shown in Figure 7.

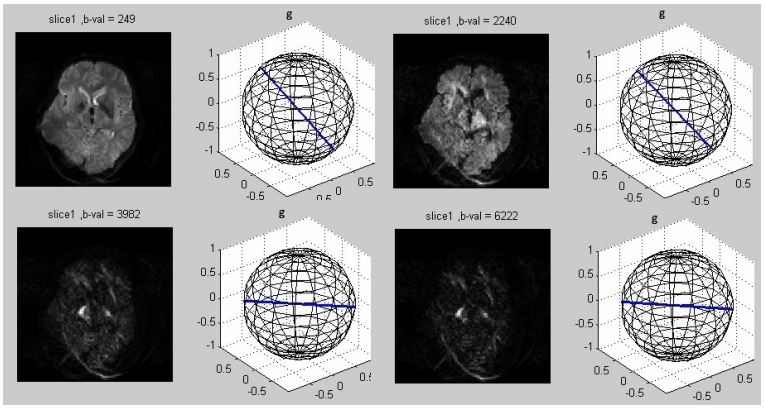


FIGURE 5 The DW-MR images of the brain with varying gradients and b-values. High b-values corresponding to heavy diffusion weights, result much noisier DW-MR images.

2.4 The MR signal intensity and Fourier transform

The signal decay in connection with Equation (2.3) can be formulated as

$$S/S_0 \ = < e^{i\phi} > = < e^{i\gamma \int_0^T G(t)x(t)dt} >, \tag{2.7}$$

where the signal phase can be expressed as a function of time T, which is

$$\phi(T) = \gamma \int_0^T x(t)G(t)dt, \tag{2.8}$$

where $x(t)$ is the path of the particle, see Tuch (2002), Stieltjes et al. (2013).
For free diffusion in the direction of the gradient, the Fourier transform is obtained by integrating out the Brownian path $x(t)$ with respect to the probability,

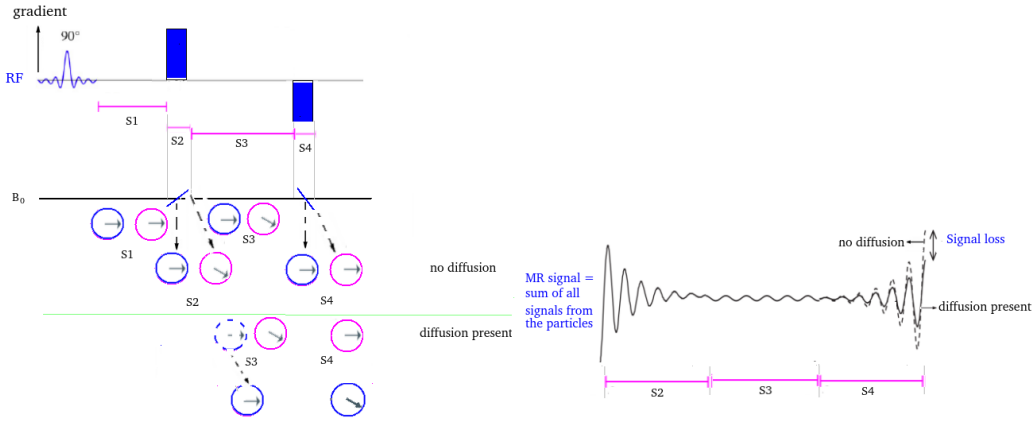


FIGURE 6 The left part of the figure depicts an example of the phase changes in two water molecules at different locations during a dephase-rephase experiment with gradient application. If there is no diffusion after the first gradient is applied, the second gradient rephases the magnetization under the condition that the strength and length of the second gradient is identical to the first as shown above the green line. When there is diffusion, for example for the molecule marked in blue under the green line, the molecules perform a random walk moving away from their initial locations. It leads to signal loss, shown in the right part of the picture. The black arrows indicate phases of the MR signals from each molecule and the MR signal is the sum of all signals from the molecules. The right plot of Figure 6 describes the possibility of the signal loss after the phase changes, see also Mori (2007).

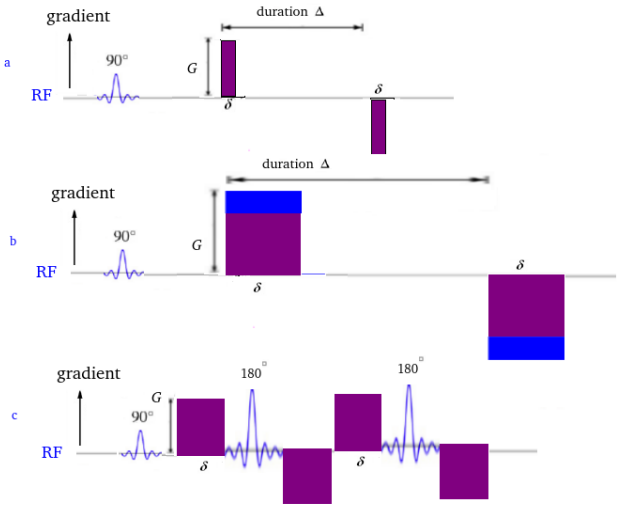


FIGURE 7 Several ways to achieve and manipulate the diffusion weighting. Figure 7a shows after 90 ° RF a weak diffusion weighting, which is generated by the very short diffusion gradients. Figure 7b demonstrates that the diffusion weighting can be manipulated by changing the values of the gradient parameters $(G, \delta, \triangle)$. Figure 7c describes spin-echo sequences with more complicated diffusion-weighting gradients.

as

$$\frac{S}{S_0} = \exp\left(-\frac{\gamma^2}{2}\int_0^T\int_0^T G(t)G(t')\langle X(t)X(t')\rangle dtdt'\right)$$
$$= \exp\left(-\gamma^2 D\int_0^T\int_0^T G(t)G(t')(t\wedge t')dtdt'\right),$$

and by using integration by parts together with the equation

$$\int_0^T G(t)dt = 0$$

we obtain the log-signal decay

$$\log\frac{S}{S_0} = -D\gamma^2\int_0^T dt\int_t^T dt' G(t)G(t')(t'-t), \tag{2.9}$$

see details for instance in Mori (2007), Stieltjes et al. (2013), where T is the total duration of the diffusion gradient.

The diffusion weighting factor (the b-value) Equation (2.9) explicitly contains a well-known formula of the b-value defined by

$$b = \gamma^2\int_0^T dt\int_t^T dt' G(t)G(t')(t'-t). \tag{2.10}$$

Hence the signal intensity can be written as a function of b-value, which is

$$S(b) = S_0\exp(-bD). \tag{2.11}$$

This simplest expression only works for the free diffusion, and hence here D is a constant. Considering a diffusion weighting sequence produced by a pair of bipolar gradients, the time caused signal dephasing can be measured by the function $\int_0^T G(t)dt$, and the amount of dephase is represented as the red area in Figure 8. The integral in Equation (2.10) can be then calculated with three time intervals recorded at time lags $t_1 = 0, t_2 = \delta, t_3 = \triangle$ and $t_4 = \delta + \triangle$, see Mori and Van Zijl (1995), Stieltjes et al. (2013), Bammer (2003), resulting in the famous Stejskal and Tanner b-value Equation, $b = \gamma^2 G^2\delta^2(\triangle - \delta/3)$, see Stejskal and Tanner (1965).

The Fourier transform The diffusion MRI is capable of capturing the mean (average) displacement of water molecules through the probability density function $P(x,t)$, which represents the sum of the diffusion from all water molecules over the microscopic level existing in the image volume element known as *voxel*. The resolution of the voxel is defined by the macroscopic spatial encoding (Tuch , 2002), referring as a measurable pixel in the brain. In the PGSE experiment (Stejskal and Tanner , 1965), the gradient duration is much shorter than the duration between two gradients, $\delta << \triangle$. The signal phase in Equation (2.3) is reformulated as $\phi = \gamma G\delta(X_t - X_0)$, where γ is the gyromagnetic ratio and $G\delta$ is the gradient scheme.

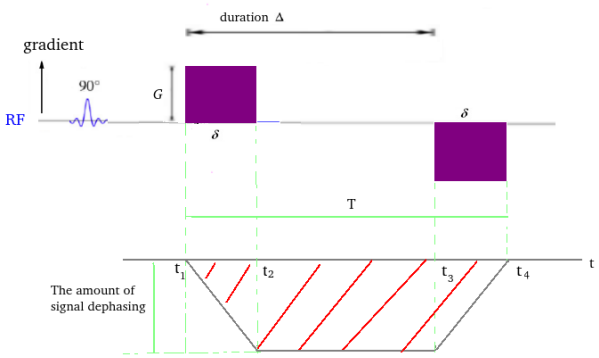


FIGURE 8 The amount of signal dephasing caused by the dephasing gradient can be represented as a function of time t, dephasing as the red area in Figure 8.

Let $\mathbf{q} = \gamma G \delta$, when the time interval is short to be ignored, the signal decay in the physical 3D world can be written as

$$S(\mathbf{q})/S_0 = \int p(x,t) \exp(i\mathbf{q}^T x) dx, \tag{2.12}$$

where S_0 is the non-diffusion weighted signal intensity at the voxel v. Hence, the signal decay is the Fourier transform of $p(x,t)$, which allows us to describe the probability distribution of diffusion from the spin echo signal intensity by the inverse Fourier transform, that is,

$$p(x,t) = S_0^{-1} \mathcal{F}^{-1}\Big(S(\mathbf{q})\Big). \tag{2.13}$$

When the displacement is Gaussian, we obtain the commonly used Stejskal-Tanner signal intensity equation. Accordingly, in the diffusion-MR experiment the water molecules at the voxel v emit a spin echo signal with amplitude

$$S_v(\mathbf{q}) = S_0 \exp\big(-\frac{1}{2}\mathbf{q}^\top \overline{D}_v \mathbf{q}\big) = S_0 \exp\big(-b\mathbf{g}^\top \overline{D}_v \mathbf{g}\big), \tag{2.14}$$

where $\overline{D}_v$ denotes diffusion coefficients or named diffusion tensor at voxel v. The gradient pulse $\mathbf{q} \in \mathbb{R}^3$ is a predefined parameter of the MR experiment. The diffusion weighting is $b = |\mathbf{q}|^2/2$, the b-value, and $\mathbf{g} = \mathbf{q}/|\mathbf{q}| \in \mathcal{S}^2$ indicates the gradients on the unit sphere. Equation (2.14) in particular shows that the probability distribution of the departure of water molecules can be implicitly represented by the covariance matrix $\overline{D}$ (for simplicity we omitted subscript v here and thereafter), and describes the diffusion tensor imaging model (DTI, Basser et al. (1993, 1994b)) when $\overline{D}$ is forming as 3×3 positive semi-definite matrix. DTI is one of key concepts in this thesis, which has also been extended to more complex situations, where a single diffusion tensor is not enough to describe the displacement distribution at all locations within a voxel, and we should rather think about a population of diffusion tensor within the same voxel. One of the solutions could be that the displacement distribution of a water molecule $(X_t - X_0)$ is modeled

as a mixed Gaussian starting from the initial location x during a unit time with Fourier transform

$$\frac{S_x(\mathbf{q})}{S_0} \;=\; E_x\Big(\exp\big(i\ \mathbf{q}^\top(X_t - X_0)\big)\Big) \;=\; \int\limits_{\mathcal{M}^+} \exp\Big(-\frac{1}{2}\mathbf{q}^\top \overline{D}^*\,\mathbf{q}\Big) dQ_x(\overline{D}^*)\;, \quad (2.15)$$

where Q_x is a probability distribution on the space of positive matrices $\mathcal{M}^+$. This equation implicitly describes that the symmetric and positive definite matrix-valued field ($\overline{D}^*$) as covariance matrices of water displacement can explain the geometry of an underlying media.

The Fourier transform gives us a path to describe the diffusion process by means of diffusion tensor $\overline{D}$. Computational accuracy and efficiency in the estimation of the tensor matrix $\overline{D}$ and the relatives are the key objectives in this thesis.

3 DIFFUSION TENSOR IMAGING AND ITS EXTENSIONS

In anisotropic media the mobility of the molecules is orientation dependent and can not be represented in terms of one single diffusion coefficient. The three-dimensional (3D) diffusion process modeled by means of simple 2nd order diffusion tensors was introduced by Basser et al. (1993, 1994a), and this well-known diffusion MRI reconstruction technique is named diffusion tensor imaging (DTI), which was considered in **PI** and **PII** of this thesis. DTI as an established approach during almost two decades provides a systematic description of diffusion anisotropy and fibre tacking in the study of the brain connectivity, which is capable of quantifying in vivo the diffusion displacement of water molecules at the microscopic level describing the structural information and the geometric organization of brain anatomy. DTI has been successfully applied in many clinical studies for detecting common brain disorders, such as stroke (Mori , 2007), intrinsic tumor and demyelination (Giussani et al. , 2010), and others, see e.g. Sundgren et al. (2004), Ringman et al. (2007), Wozniak et al. (2013). However, it suffers from an intrinsic limitation that the displacement of water molecules is assumed to follow a 3D homogeneous Gaussian distribution within each voxel (Basser , 1995). Tuch et al. (1999), Tuch (2002) developed an approach to detect complex tissues within each voxel, overcoming the limitation of DTI in the regions of heterogeneous structure which restricts diffusion. This method is termed the higher angular resolution imaging (HARDI), which is considered in **PI** and **PII**. Meanwhile, Niendorf et al. (1996) pointed out that the water diffusion in biological structures indicates non-Gaussian diffusive behavior due to the barriers of the cellular compartments and membranes inside biological tissue. In the last few years, several approaches, such as Clark et al. (2002), Özarslan and Mareci (2003), Yablonskiy et al. (2003), characterized the non-Gaussian properties of water diffusion. Jensen and Helpern (2003), Jensen et al. (2005) extended DTI and obtained significantly more promising results of characterizing the diffusion and tissue structure. This method is called diffusion kurtosis imaging (DKI), which is considered in **PIII** and **PIV**.

3.1 DTI

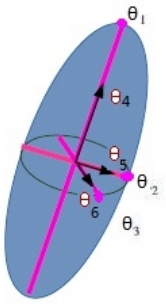
Owing to the highly directional architecture of the white matter (WM), DTI has been widely used for the study of WM integrity and changes in diffusion anisotropy through a simple second order symmetric and positive semi-definite tensor matrix $\overline{D}$ and the model gives rise to

$$S(b, \mathbf{g}) = S_0 \exp\Big(-bD\Big) = S_0 \exp(-b\mathbf{g}^T \overline{D}\mathbf{g}), \tag{3.1}$$

where b gives the weights to the diffusion, $\mathbf{g} \in \mathbb{R}^3$ contains the pulse gradients on the unit sphere and D is formally named *apparent diffusion coefficients* (ADC), $D =$ ADC, in Dong et al. (2004). In DTI, the geometric structure (morphology) of the 2nd order tensor typically is an ellipsoid determined algebraically by a vector parameter $\boldsymbol{\theta} \in \mathbb{R}^6$ containing six spatial random variables. This tensor parameter can be easily vectorized to be a 3×3 tensor matrix $\overline{D} := (D_{i,j} : 1 \leq i \leq j \leq 3)$ with at most six unique elements due the symmetry and $\theta_1 = D_{11},\ \theta_2 = D_{22},\ \theta_3 = D_{33},\ \theta_4 = D_{12},\ \theta_5 = D_{13},\ \theta_6 = D_{23}$, see the description in Figure 9. DTI then can be parametrized as

$$S = S_0 \exp\Big(Z(b, \mathbf{g})\theta\Big), \tag{3.2}$$

where Z is a $m \times 6$ design matrix and here $m = 1$.



$$\theta \leftrightarrow \overline{D} := \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{12} & D_{22} & D_{23} \\ D_{13} & D_{23} & D_{33} \end{bmatrix}$$

FIGURE 9 An anisotropic 2nd order tensor represented as an ellipsoid can be vectorized as a 3×3 symmetric positive semi-definite matrix $\overline{D}$.

3.2 Tractography

The 3×3 tensor matrix $\overline{D}$ is composed of three eigenvector-eigenvalue pairs $(\lambda_i, \mathbf{v}_i, i = 1, \cdots, 3)$ and $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$, since negative diffusion has nonphysical sense. This equation delivers detailed information on the principal direction ($\mathbf{v}_1$) of diffusion in such a way that for each Gaussian component in the mixture,

the MR signal is highest when g belongs to the eigenspace of the smallest eigenvalue of $\overline{D}$, and lowest in the principal direction.

An isotropic diffusion tensor (DT) has a ball shape, where the off-diagonal elements in $\overline{D}$ are vanished to be zero and the diagonal entries are identical, that is $D_{11} = D_{22} = D_{33}$ and $D_{11} = \lambda_1, D_{22} = \lambda_2, D_{33} = \lambda_3$. When the diagonals are not equal, they describe the diffusion coefficients along certain directions: D_{11} for x direction, D_{22} for y direction and D_{33} for z direction. Two common morphologies of the tensor are prolate when $D_{11} \geq D_{22} = D_{33}$ and oblate ($D_{11} = D_{22} \geq D_{33}$), see an example in Figure 10. A typical anisotropic 2nd order tensor as also shown in Figure 9 is nondegenerated that all eigenvalues differ and all off diagonals in $\overline{D}$ appear. Such a tensor describes the diffusion coefficients along different directions, implying the corresponding fibre has an arbitrary orientation. The principal eigenvector corresponding to the largest eigenvalue of the tensor can sometime directly indicate the fibre orientation. A scalar measure called the mean diffusivity (MD) is commonly used to describe the strength of the diffusion as the average values of diffusivity in a diffusion process, which has a formula

$$\text{MD} = (D_{11} + D_{22} + D_{33})/3. \tag{3.3}$$

The degree of anisotropy can be simply measured by the division between the

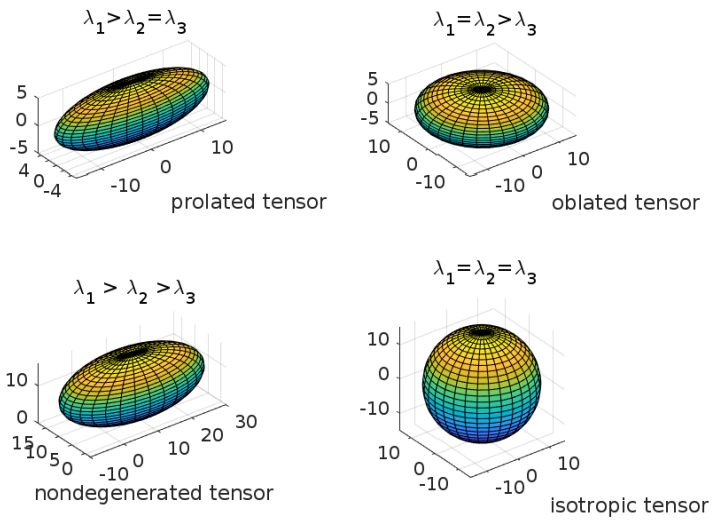


FIGURE 10 The morphologies of four DTs.

largest and smallest eigenvalues. More often people use the fractional anisotropy (FA) to express anisotropy and is defined as

$$\text{FA} = \sqrt{\frac{3}{2}} \frac{\sqrt{(D_{11} - MD)^2 + (D_{22} - MD)^2 + (D_{33} - MD)^2}}{\sqrt{D_{11}^2 + D_{22}^2 + D_{33}^2}}. \tag{3.4}$$

These quantities can be further visualized as maps (see Figure 11) to describe the properties of the diffusion in a region of the underlying object. The fibre orientation can be estimated by measuring the diffusion anisotropy and depicted

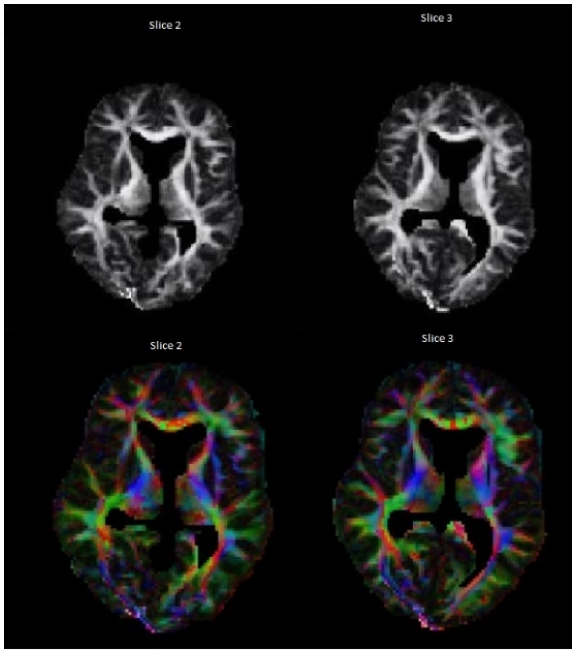


FIGURE 11 The FA maps (bottom) from two consecutive slices of a human brain encoding by two Red-Green-Blue (RGB) colormaps. The three colors correspond to x-z-y coordinates of the principal eigenvector and show the orientations of the fibres. The two MD maps (top) give an overview of the architecture of the regions of the human brain.

by the underlying connection of the tensor ellipsoid (Barmpoutis et al. , 2009b, Basser et al. , 2000, Behrens et al. , 2007, Descoteaux , 2010, Zhu et al. , 2007).

The tractography of human brain contains millions of fibres and their orientations (Tournier et al. , 2012, Özarslan et al. , 2006), consequently the tracts from a small area of the human brain can be very complicated as illustrated in Figure 12. Characterizing the morphologies of DTs hence has been used in current imaging studies for tracking fibres and for reconstructing the tractography of the human brain. Figure 13 sketches two possible fibre tracts (the dark lines) in human brain from the estimated tensor shape. It also indicates that the tensor shape in human brain can be much more complicated than an ellipsoid, which indicates that the 2nd order tensor of a low angular resolution may be insufficient to describe the complex structure of the brain. In fact, the intrinsic limitations in DTI come from the original model assumption that the water diffusion has a Gaussian distribution (Basser et al. , 2000, Barmpoutis and Vemuri , 2010).

High order diffusion tensors (HODT) Modeling diffusivity is very complicated in the real scenario as in human brain. Using classic 2nd order tensors as the tensor parameter to capture the complex feature of biological tissue such as fibre crossing may be practically infeasible, which hence results in the lost of major anatomical information. Özarslan and Mareci (2003) propose a generalized

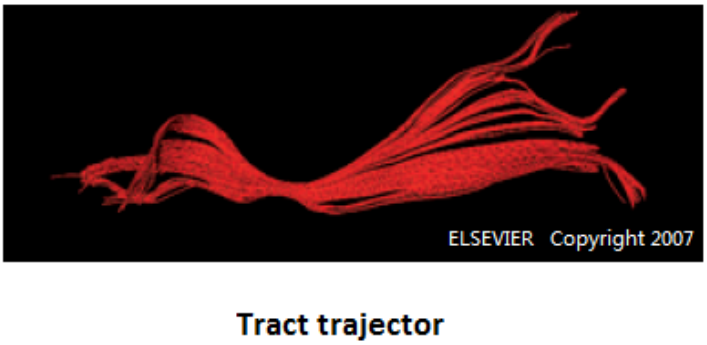


FIGURE 12 Fiber tracts in human brain. The figures are reprinted from Mori (2007). Copyright (2007), with permission from Elsevier, *http://www.elsevier.com.*

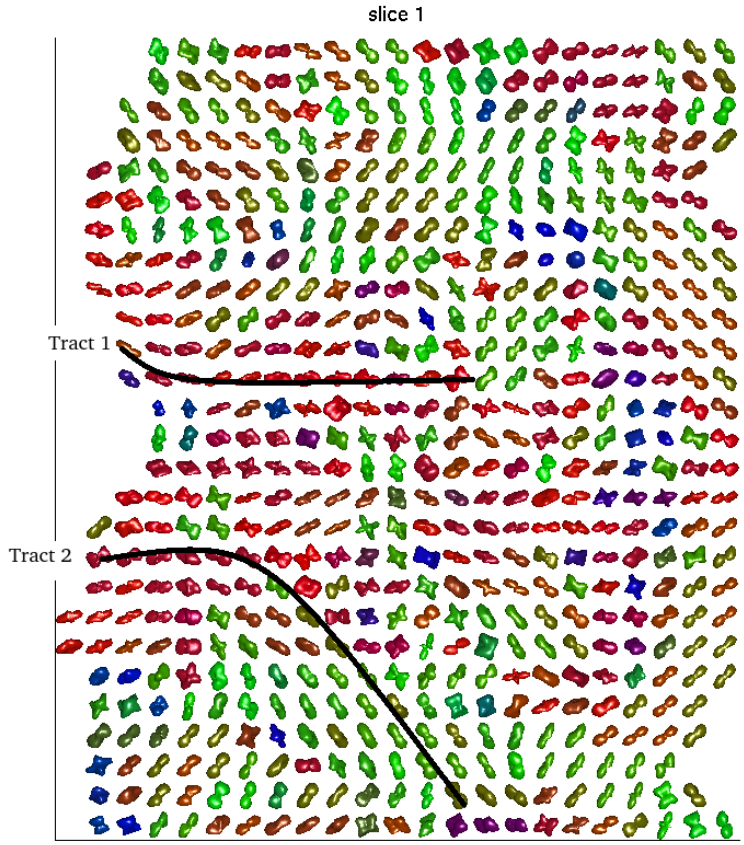


FIGURE 13 Two sketches of fibre tracts in human brain retrieved by the estimates of the tensor parameters. The colors indicate the main direction of the principal eigenvalue of the tensor: red, left-right; green, anterior-posterior; blue, superior-inferior.

Stejskal-Tanner equation to express the diffusivity as a function of gradients

$$D := \sum_{\ell_1=1}^{3} \sum_{\ell_2=1}^{3} \cdots \sum_{\ell_{2n}=1}^{3} D_{\ell_1,\ell_2,\ldots,\ell_{2n}} g_{\ell_1} g_{\ell_2} \cdots g_{\ell_{2n}} \ , \tag{3.5}$$

being homogeneous polynomials. The 4th order tensor as one of common HODTs has been widely studied in the literature (e.g. Barmpoutis et al. (2009), Barmpoutis and Vemuri (2009), Basser and Pajevic (2007)), based on the high angular resolution imaging (HARDI) acquisitions, see Tuch et al. (1999), Tuch (2002), Descoteaux (2010). A 4th order tensor can be vectorized as a 6×6 symmetric positive semi-definite matrix $\overline{D}$, see e.g. Barmpoutis et al. (2007), and typically has a cross shape as illustrated in Figure 14.

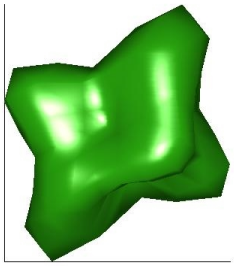


FIGURE 14 A typical shape of a 4th order tensor that is in general controlled by 15 unique tensor coefficients.

3.3 HARDI

The diffusion reflecting the displacement distribution of water molecules may exhibit non-Gaussian behavior due to the disturbance from the structure of biological tissues (Cory , 1990). Higher angular resolution imaging (HARDI) developed by Tuch et al. (1999), Tuch (2002), refers to a sampling technique that originally intends to discover evidence of spatially non-Gaussian diffusion in the white matter of human brain. Accordingly, using the HARDI data to model ADC is possible to overcome the limitations in DTI, especially for approximating the complex tissue geometry. Figure 15 reflects the non-Gaussian behavior in the human brain with multi-lobed diffusivity profiles. This modality has been extended later on from a signal-shell acquisition to the multiple-shell scheme, allowing using multiple b-values to acquire data in the sampling procedure.

3.4 DKI

The Gaussian assumption of the water diffusion is argued to diverge significantly from the genuine in biological tissue. It is known that appendages are highly

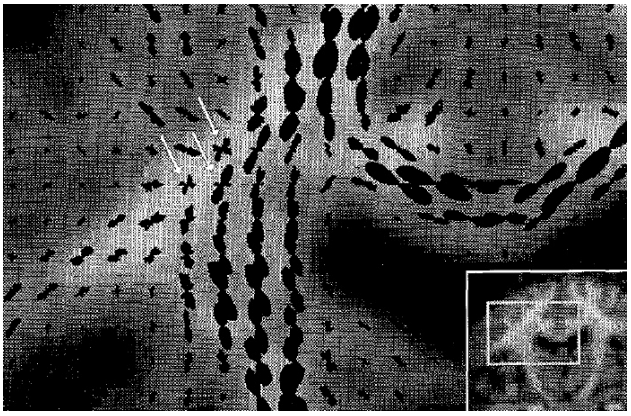


FIGURE 15 The non-Gaussian diffusion at higher angular resolution (mainly refers to the 4th order tensors marked by the arrows) is discovered at the corona radiate lateral to the lateral ventricle and medical to the Sylvian fissure in human brain, with zoom. The figure is adapted from Tuch et al. (1999), with permission from ISMRM.

complex and rich structures, consisting of multiple biological tissues. In the human brain such tissues include for example cell membranes, boundaries and other complex compartments, where water diffusion is far away from Gaussian. Diffusion kurtosis imaging (DKI) proposed by Jensen et al. (2005) as an extension of DTI has recently become popular in quantifying the degree of diffusional deviation from Gaussianity. The model is addressed to be useful in diagnosis of brain disorder, such as Alzheimer's disease (Lu et al. , 2006b) and ischemic stroke (Helpern et al. , 2009).

The DKI model is derived through the Taylor expansion of signal decay truncated at order 4 and may be expressed by

$$S(b,\mathbf{g})/S_0 = \exp(-bD + \frac{1}{6}b^2D(g)^2K(\mathbf{g})) \tag{3.6}$$
$$= \exp\bigg(-b\sum_{\ell_1,\ell_2=1}^{3} g_{\ell_1}g_{\ell_2}D_{\ell_1,\ell_2} + \frac{b^2}{6}(\sum_{\ell_1=1}^{3}\frac{D_{\ell_1\ell_1}}{3})^2 \sum_{\ell_1,\ell_2,\ell_3,\ell_4=1}^{3} g_{\ell_1}g_{\ell_2}g_{\ell_3}g_{\ell_4}W_{\ell_1,\ell_2,\ell_3,\ell_4}\bigg),$$

as in Jensen et al. (2005), Ghosh et al. (2014), where the signal decay is defined as a function of the b-value and $K(\mathbf{g})$ indicates the apparent kurtosis coefficient. There are two unknown components in the model, the 2nd order diffusion tensor $\overline{D}$ and the 4th order kurtosis tensor $\overline{W}$. Additionally, the model requires at least three distinct b-values ($\leq 3000s/mm^2$) and fifteen distinct gradient acquisitions. All the three constraints arise a nonlinear regression problem including a nonlinear constraint in the estimation of the tensor parameters. In comparison with MD and FA in DTI, mean kurtosis (MK) and kurtosis anisotropic (KA) are two primarily interesting metrics for DKI.

MK and KA are formulated as

$$MK = \frac{1}{4\pi}\int d\mathbb{O}_{\mathbf{g}}K(\mathbf{g}) \text{ and } KA = \sqrt{\frac{1}{4\pi}\int d\mathbb{O}_{\mathbf{g}}(K(\mathbf{g}) - MK)^2}, \tag{3.7}$$

where $\mathbb{O}$ denotes the unit sphere and $\mathbf{g} \in \mathbb{O}$, see Jensen and Joseph (2010), Poot et al. (2010), Tabesh et al. (2011).

3.5 Other related models

This thesis also considered some other signal models. In **PIII** we had talked about biexponential model and multicompartment model in the simulation studies, here we just give a short description.

Biexponential model Several studies have shown that signal decay in the brain exhibits non-monoexponential diffusion and can be approximated well by using biexponential curves at very high b-values (5000 s/mm^2), see Maier et al. (2004), Jensen et al. (2005), Maier and Mulkern (2008). The biexponential model is represented by

$$S/S_0 = f\exp(-bD_f) + (1-f)\exp(-bD_s), \tag{3.8}$$

where D_f and D_s are the diffusion coefficients of a fast and a slow diffusion component, respectively, corresponding to the water fraction f and $1-f$, and is computed by means of the nonlinear least-squares Levenberg-Marquardt algorithm. This model is suitable for studying water diffusion both in the gray matter (GM) and in the white matter (WM) of the brain, and manages to reveal the diffusional difference between GM and WM from the high b-value diffusion MRI data. Therefore, we use this model as a reference model in the simulation studies to test the results under other imaging protocols. The reader should bear in mind that, although a simulation study is an inverse process of estimation, choosing appropriate pulse gradients is challenging in terms of making synthetic data to be close to the real ensemble. This is in connection with the experimental design, an important topic in statistics.

Multicompartment model The multicompartment model (Tuch et al. , 2003, Behrens et al. , 2003, 2007) is used to infer multiple fibre orientations in each voxel. The model may be expressed by

$$S_i/S_0 = (1-f)\exp(-b_i d) + f\exp\Big(-b_i d g_i^T \overline{D}(\vartheta,\varphi) g_i\Big) \tag{3.9}$$

as in Behrens et al. (2003), where f and $1-f$ are the water fractions, and d denotes the diffusivity. The anisotropic DT, D, is along the fibre direction (ϑ,φ).

A specific case described by Zhu et al. (2013), also called "ball-and-sticks" model,

$$S_i/S_0 = f_0\exp(-b\theta_0) + \sum_{j=1}^{M}\exp\Big(-b\theta_j(g_i^T u_j)\Big), \tag{3.10}$$

is used for tracking complex fibre orientations in WM. The model contains three vector parameters including the diffusion coefficients $\theta_0 \cdots, \theta_M$, the water fractions $f_0, \cdots, f_M$, and the WM fibre orientations $u_1, \cdots, u_M$, where M is the maximum number of anisotropic compartments.

3.6 Positivity

The diffusivity is in DW-MRI a real valued positive function, implying the tensor of any order constrained by positivity. Numerous works (Barmpoutis et al. , 2009, Barmpoutis and Vemuri , 2010, Qi et al. , 2010) have been dedicated to the study of the positivity constraint in DT. The typical idea originates from Hilbert's Theorem (Hilbert , 1888) that any real-valued positive function can be written as a sum of three squares of quadratic forms. Barmpoutis and Vemuri (2010) pointed out that for any Kth order tensor, the diffusivity can be modeled by a sum of squares of $\frac{K}{2}$th-order homogeneous polynomials

$$D = \sum_{j=1}^{M} \text{poly}(\mathbf{g}; c_j)^2$$

with $M \leq \frac{(2+K/2)!}{2(K/2)!}$, and there exists a map between the polynomial coefficients $\mathbf{c}$ and tensor parameter $\boldsymbol{\theta}$. For the 4th order tensor (HOT4) the diffusivity function may be expressed by

$$D = \sum_{i=1}^{3} (\mathbf{g}^T C_i \mathbf{g})^2 \tag{3.11}$$

(Papadopoulo et al. , 2014), where C_i is a symmetric 3×3 matrix. Here K must be an even number since an odd order of tensor has no physical meaning.

Estimating tensor coefficients The signal model in general transforms water diffusion into diffusion tensor and the pre-set gradients. When the tensor is estimated, tractography becomes feasible through the classification of the morphologies of the tensors. How to obtain accurate and efficient tensor estimates under certain signal model hence becomes a statistical problem. We use the simplest model DTI to describe the problem. When signal S is observable from the noisy-free DW-MR images, the tensor parameter can be calculated accurately from Equation (3.2). This, however, is not true in general and the DW-MR images contain various noise components affecting DTs. In the following chapter we will describe the MR signal measurement that we observed from the DW-MR images, the random noise from the images and the commonly used methods for tensor estimation.

4 THE MR SIGNAL MEASUREMENT, RANDOM NOISE AND MLE

4.1 Signal measurement in MRI

The MR signal measurement y is generated from the real and imaginary parts of the MR images, where the signal intensity S is corrupted by the noise ε, resulting in $y = |S + \varepsilon| = \sqrt{(S + \varepsilon_r)^2 + \varepsilon_i^2}$. Due to the nonlinear mapping, the noise distribution is no longer Gaussian (Gudbjartsson and Patz , 1995), but is a complex-valued variable. If $(S + \varepsilon_r)$ and ε_i are two independent Gaussian random variables, then the probability distribution of the MR signal measurement will be Rician (Rice , 1944, Henkelman , 1985, Bernstein et al. , 1989, Andersen , 1996) and has a Rician density function

$$p_{S,\sigma^2}(y) = \frac{y}{\sigma^2} \exp\left(-\frac{y^2 + S^2}{2\sigma^2}\right) I_0\left(\frac{yS}{\sigma^2}\right) \mathbb{1}(S \geq 0), \tag{4.1}$$

where $I_0 = \frac{1}{\pi} \int_0^\pi \exp(z \cos t) dt$ is called the modified zeroth order Bessel function of the first kind, which is a special case of the α-order modified Bessel function of the first kind ($I_\alpha(\cdot)$) with a general expression

$$I_\alpha(z) = \sum_{n=0}^{\infty} \frac{(z/2)^{2n+\alpha}}{n!(n+\alpha)!} = \frac{1}{\pi} \int_0^\pi \exp(z \cos t) \cos(\alpha t) dt, \quad \text{for } \alpha \in \mathbb{N},$$

and σ^2 is the variance from the noise components. We use $\mathbb{1}(\cdot)$ as the indicator function to emphasize the signal intensity S is a real-valued positive quantity. If we only consider the random noise in the MR images and ignore the structured noise caused by for example bulk movement or blood flow (Mori and Van Zijl , 1995), then the real and imaginary noise components ε_r and ε_i form the complex-valued noise, are independent and Gaussian distributed with zero mean and common variance σ^2 (Henkelman , 1985, Koay and Basser , 2006, Zhu

et al. , 2007) and have a joint density

$$p_{S,\sigma^2}(\varepsilon_r,\ \varepsilon_i) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\varepsilon_r^2+\varepsilon_i^2}{2\sigma^2}\right). \tag{4.2}$$

The noise ε is a complex-valued Gaussian noise, and in convention is called "Rician noise". This is because the noise term couples the MR signal measurement and itself *does not* have the Rician distribution.

Multichannel signal model The MR signal measurement in multichannel case is straightforwardly along the following expressions. Suppose we have κ channels, $\kappa > 1$, the signal intensity again is corrupted by a complex-valued noise $\varepsilon^{(\ell)}$

$$x = \frac{1}{\kappa}\sum_{\ell=1}^{\kappa}|S+\varepsilon^{(\ell)}|^2 \tag{4.3}$$

in forming the MR images, where the data $y = \sqrt{x}$. If the complex noise $\varepsilon^{(\ell)} = \varepsilon_x^{(\ell)} + i\varepsilon_y^{(\ell)}$ from each channel is independently generated by two $i.i.d.$ Gaussian random variables $\varepsilon_x^{(\ell)}, \varepsilon_y^{(\ell)} \sim \mathcal{N}(0, \kappa\sigma^2)$, $\ell = 1, \ldots, \kappa$, with noise parameter σ^2, the MR signal measurement follows the Rician distribution with density

$$p_{S,\sigma^2,\kappa}(y) = \frac{S}{\sigma^2}\left(\frac{y}{S}\right)^{\kappa}\exp\left(-\frac{y^2+S^2}{2\sigma^2}\right)I_{\kappa-1}\left(\frac{yS}{\sigma^2}\right), \tag{4.4}$$

see Aja-Fernández and Vegas-Sanchez-Ferrero (2004). If the random variable of the observation (data) $Y = 0$, we have then $\varepsilon_x^{(\ell)} = -S,\ \varepsilon_y^{(\ell)} = 0, \ell = 1, \ldots, \kappa$, with likelihood contribution

$$(2\pi\kappa\sigma^2)^{-\kappa}\exp\big(-\frac{S^2}{2\sigma^2}\big)\ .$$

Specially, when $\kappa = 1$ we are back in the simple case as in Equation (4.1).

Let the MR signal in DTI be $S = \exp(Z\boldsymbol{\theta})$, so that $\theta_0 \in \boldsymbol{\theta} \in \mathbb{R}^7$ stands for the signal without diffusion weighting S_0 and the dimension of the corresponding design matrix will be $m \times 7$. When we only consider the real random noise in the DW-MR images, the logarithmic signal measurement then can be linearized by log transformation and falls into the linear framework. A log-linear model hence gives rise to

$$\log y|Z, \boldsymbol{\theta} = Z\boldsymbol{\theta} + \exp(-Z\boldsymbol{\theta})\varepsilon, \tag{4.5}$$

where we use the fact that $\log(1+\exp(-Z\boldsymbol{\theta})\varepsilon) \approx \exp(-Z\boldsymbol{\theta})\varepsilon$. The error ε in different DTs can have different distributions and heterogenous variance. When it follows a Gaussian distribution, the likelihood of the MR signal measurement reduces to be Gaussian with a common assumption of zero mean. In fact, the logarithm of the MR signal measurement in Equation (4.1) starts to approximate the Gaussian distribution with mean $\mathbb{E}(\log Y_i|Z, \boldsymbol{\theta}) = Z\boldsymbol{\theta}$ and standard deviation

$1/\text{SNR}$ (SNR $:= S/\sigma$), when the value of SNR is equal to 3, and the approximation works well with SNR > 5, see Gudbjartsson and Patz (1995), Salvador et al. (2004), Zhu et al. (2007).

Among a number of different estimation approaches, the most common and standard methods for linear regression are linear least squares (LLS) and weighted least squares (WLS), which are usually applied with the Gaussian models. When the noise has a common finite variance, LLS is applied; whereas when ε has violated variance, WLS is the right choice.

4.2 LLS and WLS

The theory behind LLS is minimizing the error between the observed data $\log Y$ and the predicted data $\log \hat{Y} = \mathbb{E}(\log Y | Z, \hat{\boldsymbol{\theta}})$ measured in terms of their Euclidean norm, known as the sum of square residuals function,

$$f(\boldsymbol{\theta}) = ||\log Y - Z\boldsymbol{\theta}||_2^2. \tag{4.6}$$

If the number of observed signal measurements $m >$ dim$(\boldsymbol{\theta}) = d + 1$, a unique solution $\widehat{\boldsymbol{\theta}} = (Z^T Z)^{-1} Z^T \log \mathbf{y}$ can be obtained, which is called the original least squares (OLS) estimator. The variance of the error can be estimated by $\widehat{\sigma}^2 = \frac{(\mathbf{Y} - Z\widehat{\boldsymbol{\theta}})^T (\mathbf{Y} - Z\widehat{\boldsymbol{\theta}})}{m-d}$, see e.g. Patterson and Thompson (1971), where $m - d$ in statistical convention is called the number of degree of freedom.

When the error on different acquisition has violated variance, Equation (4.5) becomes a heteroscedastic linear model (Goldberger , 1964). WLS then is the method to apply, which minimizes the weighted sum of squared residuals

$$f(\theta) = \sum_{i=1}^{m} w_i (\log Y_i - Z_i \boldsymbol{\theta})^2. \tag{4.7}$$

When the weights $\boldsymbol{w}$ equal to the inverse model variance (square of SNR), $\boldsymbol{w} = \exp(2Z\boldsymbol{\theta})\sigma^{-2}$, $\widehat{\boldsymbol{\theta}}$ is the best linear unbiased estimator (BLUE), see Aitken (1935). Since σ^2 is unknown, in practice the choice of the weights can be for example $\boldsymbol{w} = Y$ or $\boldsymbol{w} = \exp(Z\boldsymbol{\theta})$, see **PIII** for details.

4.3 MLE

In the standard linear model, the response typically follows an implicit assumption that the components are $i.i.d.$ Gaussian random variables which have a roughly linear relation to the explanatory variables. Such an assumption does not hold for the Rician model of the MR data, especially when SNR is lower than 3. A linear model, when fitted to data that do not follow such a linearity assumption, may result in a bad performance for estimation and prediction. Hence, estimation

of the tensor and variance parameter with the Rician likelihood rises a nonlinear regression problem.

Nonlinear regression is a form among regression techniques, in which the observations are modeled as a function of a nonlinear combination of the model parameters and one or more independent explanatory variables. The maximum likelihood estimation (MLE) is among the most general and popular estimation techniques and can directly solve the nonlinear regression problem. It usually works with the log of the likelihood function utilizing the monotonic property of the logarithm function and provides estimators by maximizing the object function: In DTI, the object function is the log-likelihood of the MR signal measurements $Y = (Y_i, i = 1, \cdots, m)^T$, which is given by

$$\log L(\beta|Y, Z) = \text{const} - m\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(Y_i^2 + \exp(2Z_i\boldsymbol{\theta})\right) + \sum_{i=1}^{m}\log I_0\left(\frac{Y_i\exp(Z_i\boldsymbol{\theta})}{\sigma^2}\right). \tag{4.8}$$

The parameter vector is specified by $\beta = \{\boldsymbol{\theta}, \sigma^2\}$, "const" denotes the constant term in short and L is also called the loss function in decision theory. The primary attraction of the maximum likelihood estimator (MLE) is in its asymptotic properties, consistency and asymptotic normality (under regularity conditions). Accordingly, the sampling distribution of $\hat{\beta}_{\text{mle}}$ can be approximated by the normal distribution with mean β and covariance matrix $\mathbb{I}^{-1}(\beta)$ where the latter is the inverse of the expected Hessian matrix (Fisher information matrix). The proofs can be found in e.g. Cramér (1946), Amemiya (1985).

Under the principle of maximizing the log-likelihood, the MLEs are typically rendered by calculating the first derivative (the score) of the log-likelihood, considered as a function of each unknown parameter, set the score to zero, and then solve the system of the equations. In using numerical methods, MLEs usually cannot be expressed in closed form and the system of the equations has to be solved iteratively. The Newton-type algorithm is commonly applied in calculation of MLEs, obtaining from the asymptotic normality of MLEs. When the likelihood function is a linear model with normal noise, then the MLEs are equivalent to the ordinary least squares estimators.

4.4 The Newton-Raphson Method and Fisher Scoring

The standard Newton-Raphson iteration, also called the Newton method, is a frequently used optimization method in statistics. In line with the Newton scheme the tensor parameter $\boldsymbol{\theta}$ can be updated by

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbb{S}(\boldsymbol{\theta})H(\boldsymbol{\theta})^{-1}.$$

The idea is to use the first two terms of the Taylor series expansion, which are the score $\mathbb{S}(\boldsymbol{\theta}) := \partial Q/\partial\boldsymbol{\theta}$ and the Hessian matrix $H := \partial^2 Q/\partial\boldsymbol{\theta}^T\partial\boldsymbol{\theta}$, being a log-

likelihood quadratic in $\boldsymbol{\theta}$, to find the mode of the tensor by maximizing the log-likelihood function $Q(\boldsymbol{\theta})$, where the restriction on the score $\mathbb{S} = 0$ is running through the whole iteration to convergence. The advantage of the Newton method is that it is invariant under constant, nonsingular and linear transformations. However, calculating the second derivatives (H) may sometimes be very difficult and monitoring the convergence may not be straightforward.

In 1925, Fisher (Fisher , 1925) used the Newton method, replacing the negative Hessian matrix by its expectation, $H(\boldsymbol{\theta}) \to \mathbb{I}(\boldsymbol{\theta}) := \mathbb{E}(-H(\boldsymbol{\theta}))$.

$$\mathbb{E}_{\boldsymbol{\theta}}\Big(H(\boldsymbol{\theta})\Big) = \mathbb{E}_{\boldsymbol{\theta}}\Big(\mathbb{S}(\boldsymbol{\theta})^T\mathbb{S}(\boldsymbol{\theta})\Big)$$

is much easier, which allows us to avoid the computation of the second derivatives of the log-likelihood. In DTI we use the Cholesky decomposition on the tensor matrix $\overline{D} = LL^T$ to guarantee the symmetric and positive semi-definite, and consider $\boldsymbol{\theta}$ as a function of L. The second derivatives can usually be approximated by $\frac{\partial^2 Q}{\partial L^T \partial L} = -\sum \frac{\partial Q}{\partial \boldsymbol{\theta}_j}\frac{\partial^2 \boldsymbol{\theta}_j}{\partial L^T \partial L} - \left(\frac{\partial \boldsymbol{\theta}}{\partial L}\right)^T \left(\frac{\partial^2 Q}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}}\right)\left(\frac{\partial \boldsymbol{\theta}}{\partial L}\right)$, see e.g. Koay et al. (2006). Using the Fisher scoring, we can thus reduce the computation substantially by taking

$$\mathbb{E}(\partial Q/\partial \boldsymbol{\theta}) = 0 \qquad \text{and} \qquad \mathbb{E}\left(\frac{\partial^2 Q}{\partial L^T \partial L}\right) = \mathbb{E}\left((\frac{\partial \boldsymbol{\theta}}{\partial L})^T(\frac{\partial^2 Q}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}})\frac{\partial \boldsymbol{\theta}}{\partial L}\right),$$

as in Green (1984).

4.5 Additional robustness of Fisher scoring

Since these Newton-type methods implicitly use the quadratic assumption on the objective function, they usually converge fast. However, the algorithms are sometimes sensitive to the initial points, and hence modifications of the scoring method are necessary in order to ensure the desired increase of the likelihood in each update. A trust-region approach is among the modifications where a step parameter $\alpha \in [0, 1]$ is included and the update of $\boldsymbol{\theta}$ is then given by

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \mathbb{S}(\boldsymbol{\theta})\mathbb{I}(\boldsymbol{\theta})^{(-1)}. \tag{4.9}$$

Other serious problems in the scoring method concerning the convergence behavior may be: 1) Singular Fisher information appears in the iteration, and as a result the algorithm goes uphill. 2) Inverse Fisher information reaches a large value. To solve these problems and to make the algorithm stable, we have applied the Levenberg-Marquardt (LM) method, a combination scheme of the gradient descent method and the Fisher-scoring method to update $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \left[\mathbb{I}(\boldsymbol{\theta}) + \gamma \operatorname{diag}\Big(\mathbb{I}(\boldsymbol{\theta})\Big)\right]^{-1} \mathbb{S}(\boldsymbol{\theta}),$$

demonstrated by Marquardt (1963), where γ is the well-known LM parameter updated in each iteration to avoid a singular or ill-conditioned $\mathbb{I}(\cdot)$, and diag$(\cdot)$ is the diagonalizing operator. Optimal choices of γ have been studied in many works, and recently Fan and Yuan (2001), Yamashita and Fukushima (2001) show that $\|\mathbb{S}(\boldsymbol{\theta})\|$ and $\mathbb{S}(\boldsymbol{\theta})^T\mathbb{S}(\boldsymbol{\theta})$ perform well in practice. The alternative schemes include

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\left[\mathbb{I}(\boldsymbol{\theta}) + \|\mathbb{S}(\boldsymbol{\theta})\| \operatorname{diag}\left(\mathbb{I}(\boldsymbol{\theta})\right)\right]^{-1}\mathbb{S}(\boldsymbol{\theta}), \qquad \text{see } \textbf{PIII} \text{ and } \textbf{PIV}, \tag{4.10}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\left[\mathbb{I}(\boldsymbol{\theta}) + \mathbb{S}(\boldsymbol{\theta})^T\mathbb{S}(\boldsymbol{\theta}) \cdot I\right]^{-1}\mathbb{S}(\boldsymbol{\theta}), \qquad \text{see } \textbf{PI} \text{ and } \textbf{PII},$$

combining the trust-region and the LM schemes, where I is the identity matrix. Additionally, to reduce the computational burden, the LM scheme can be reduced to Equation (4.9) by the pre-evaluation of $\mathbb{I}(\cdot)$.

4.6 The Barrier method

The modified Fisher scoring method can solve nonlinear optimization problems robustly with fast convergent rate. When the probabilistic model contains nonlinear constraints, for example in DKI, both the diffusion tensor $\overline{D}$ and the apparent kurtosis coefficients W are constrained (see details in **PIII**), it is possible to call for some optimization tools in the scoring method to impose the constraints. The barrier method is among the popular ones. The idea is to build a "barrier "close to the boundary of the cone $\mathbb{R}^m_+$, avoiding the constraint being close to the boundary. When the barrier is close to zero, the solution of the optimization problem with the barrier will be approximated to that of the original (Ruszczyński , 2006) [Chapter 6].

For example, suppose we have a maximization (some time minimization) problem with nonlinear constraint $g_j(\tilde{\theta}) \leq 0$ and nonlinear log-likelihood function $Q(\tilde{\theta})$, after introducing the barrier, an approximated system is given by

$$\begin{aligned} &\text{maximize} && Q(\tilde{\theta}) + \mu\sum_{j=1}^{m}\ln(\nu_j) \\ &\text{subject to} && g_j(\tilde{\theta}) + \nu_j = 0, \qquad j = 1, \cdots, m. \qquad \nu_j \geq 0, \qquad g_j(\tilde{\theta}) \leq 0, \end{aligned} \tag{4.11}$$

where μ is a positive scalar called barrier parameter, which should be decreasing at each iteration. In DKI , $\tilde{\theta}$ represents all the twenty one diffusion and kurtosis tensor coefficients. Applications of the Barrier method in DKI can be found in **PIII** and **PIV**. Fisher scoring method can be applied in solving the problem described in Equation (4.11). The reader however should bear in mind that in order to implement the constrained Fisher scoring method efficiently, experienced collaboration within the regularization (modified) scheme, the choices of step parameters and the barrier parameters are required.

4.7 Generalized linear models

Solving the score functions for achieving the MLEs in DTI is obviously computationally demanding. In some cases, a nonlinear regression problem can be reduced to a linear domain by suitable transformation, for which the model becomes easy to understand. In statistics, a nonlinear regression problem is most conveniently framed in the context of generalized linear models (GLMs) for which a simple transformation can not achieve the linearity assumption in the model. GLMs are extensions of the linear models, describing the linear relation between a link function $g(\cdot)$ and the linear predictor ξ (Nelder and Baker , 1972), and the choice of $g(\cdot)$ is of a wide range but depends on the distribution of the response. In DTI, if the likelihood of the MR signal measurement can be reduced to a GLM by

$$g\Big(\mathbb{E}(Y_i|Z,\boldsymbol{\theta})\Big) = Z_{i0}\theta_1 + Z_{i1}\theta_2 + \cdots + Z_i\theta_6, \qquad \text{for} \qquad i = 1, \cdots, m, \tag{4.12}$$

where $Z\boldsymbol{\theta}$ is the linear predictor and $E(Y|Z,\boldsymbol{\theta})$ is the conditional expectation of the responses, the model will simplify the estimation scheme and consequently reduce the computation. For example, achieving the MLEs is usually fairly easy, see Darmois (1935), Pitman (1936). GLMs are quite flexible in reducing the complexity in a variety of nonlinear regression problems. However, there is an assumption that in GLMs the response y is conditionally independent and from a simple exponential family [1], and not all the density functions satisfy the assumption that the response has an exponential family density. The Rician likelihood of the MR signal measurement is such a case. In order to facilitate the nonlinear regression problem in diffusion MRI into GLMs framework and to ease the problem, we invoke the tool of data augmentation.

[1] $f_{\xi,\phi}(y) = \exp\Big(\frac{y\xi - a(\xi)}{\phi} + c(y,\phi)\Big)$, where ξ is called the natural parameter, ϕ is the scalar or dispersion parameter describing the variance of the response and $a(\cdot)$, and $c(\cdot)$ are two specific functions. In general, we have $\mu = a^{'}(\cdot)$, and $\sigma^2 = \phi a^{''}(\cdot)$, see McCullagh and Nelder (1989), Tutz (2011), Gelman et al. (2014).

PART II

5 DATA AUGMENTATION AND EM-MLE

We develop two new ideas, employing a widely used statistical tool, *data augmentation* (DA), for reducing the nonlinear regression model of the MR signal measurement into a generalized linear modeling (GLM) framework. These original contributions involved in this thesis are the first to our knowledge applied in diffusion MRI.

DA is commonly used in Bayesian statistics for bypassing difficulties in the computation of the posterior distributions. The essential idea of DA, however, arises naturally from the comprehensive statistical topic of missing data models of the general form that the likelihood can be written as

$$g(y|\beta) = \int_n f(y, n|\beta)\, dn,$$

see Tanner and Wong (1987), Robert and Casella (2004), Gelman et al. (2014). The model contains missing data, which are denoted by N. The likelihood $f(y, n|\beta)$ is called the *complete-data* likelihood and it is assumed that it is easier to deal with than $g(y|\beta)$. In DA we ease the problem by augmenting the quantity n into the observations y. Such kind of augmented data may be thought to contain missing data, although being a user-defined instrumental variable. The augmented variables are named latent or auxiliary variables in the constructed model.

5.1 DA in diffusion MRI

DA in count data space. Consider random variables (N, X), where N is Poisson distributed with mean $t > 0$, and given N, X follows the conditional distribution Gamma$(N + 1, 1/(2\sigma^2))$, which distributional assumptions result in the join dis-

tribution

$$P_{t,\sigma^2}(N=n, X\in dx) = P_t(N=n)P_{\sigma^2}(X\in dx|N=n) = \frac{(tx)^n}{(n!)^2(2\sigma^2)^{n+1}}\exp\left(-t-\frac{x}{2\sigma^2}\right)dx\ . \tag{5.1}$$

Then the following distributional results follow:

1. It is well known that the marginal density of $Y := \sqrt{X}$ is Rician with probability distribution

$$P_{t,\sigma^2}(Y\in dy) = \frac{y}{\sigma^2}\exp\left(-t-\frac{y^2}{2\sigma^2}\right)I_0\left(\frac{y}{\sigma}\sqrt{2t}\right)dy.$$

   If $t = S^2/(2\sigma^2)$, then the density is in coincidence with Equation (4.1).
2. The conditional distribution of N given Y is

$$P_{t,\sigma^2}(N=n|Y=y) = I_0\left(\frac{y}{\sigma}\sqrt{2t}\right)^{-1}\left(\frac{y^2t}{2\sigma^2}\right)^n(n!)^{-2}. \tag{5.2}$$

   In particular, we have $P_{t,\sigma^2}(N=0|Y=0)=1$.

Let $\mu > 0$, we consider two $i.i.d.$ random variables N and $N^{'}$ with Poisson(μ) distribution and define a new probability distribution

$$p_\mu(n) := P_\mu(N=n|N=N^{'}) = I_0(2\mu)^{-1}\frac{\mu^{2n}}{(n!)^2}, \qquad n\in\mathbb{N}. \tag{5.3}$$

The scaling factor $I_0(2\mu)$ is a consequence of the identity

$$I_0(2\mu) = {}_0F_1(1,\mu^2) = \sum_{n=0}^{\infty}\frac{\mu^{2n}}{(n!)^2}, \tag{5.4}$$

where ${}_0F_1(1,z)$ is the Gaussian hypergeometric function, see Gradshteyn and Ryzhik (2007). We call $p_\mu(n)$ *the reinforced Poisson distribution* with parameter μ. Let $\mu = (yS)/(2\sigma^2)$, we get Equation (5.2), and this distribution can be used to generate the augmented data N.

Let $\beta = \{\boldsymbol{\theta}, \sigma^2\}$ as in Chapter 4.3, for each data point y we augment n, which is unobserved, and then the original nonlinear regression model will be transferred to a model with the complete-likelihood

$$L\Big(t(\beta)|x(y),n,Z\Big) = \frac{(tx)^n}{(n!)^2(2\sigma^2)^{n+1}}\exp\left(-t-\frac{x}{2\sigma^2}\right). \tag{5.5}$$

As we interpreted before, this complete-likelihood comprises two parts: one is from the Gamma distribution which does not depend on the parameter $\boldsymbol{\theta}$, and the other one is from the Poisson distribution. When the tensor $\boldsymbol{\theta}$ is considered as a parameter vector, explicitly the other parameters are assumed to be known and considered as constants. Then the complete-likelihood actually is reduced into a GLM framework with the Poisson response N by omitting the Gamma part. The Poisson likelihood is standard in GLM, and here we obtain the model

$$g(\mathbb{E}(N|\text{others})) = \log(2\sigma^2t)/2 = Z\boldsymbol{\theta},$$

where the mean of the response t is $\mathbb{E}(N|\text{others}) = a^{'}(\xi)$ and $S = \exp(Z\boldsymbol{\theta})$.

DA in phase data space The signal measurement in Equation (4.1) can also be represented in the phase data space through a transformation from the real and the imaginary images to the arctangent of their ratio, see Henkelman (1985), Zhu et al. (2007), Mori (2007).

Let φ be the phase data

$$\varphi := \arg\Big(S + \varepsilon_1 + i\varepsilon_2\Big) \in [0, 2\pi)$$

such that

$$S + \varepsilon_1 = Y\cos(\varphi), \quad \varepsilon_2 = Y\sin(\varphi).$$

It follows from the chain rule that the joint density of φ and Y for fixed S and σ^2 is given by

$$\begin{aligned} p_{S,\sigma^2}(y,\varphi) &= \frac{y}{2\pi\sigma^2}\exp\Big(-\frac{1}{2\sigma^2}\big(y\cos(\varphi) - S\big)^2 - \frac{1}{2\sigma^2}y^2\sin(\varphi)^2\Big) \\ &= \frac{y}{2\pi\sigma^2}\exp\Big(-\frac{1}{2\sigma^2}\big(y^2 + S^2 - 2Sy\cos(\varphi)\big)\Big) \\ &= p_{S,\sigma^2}(y)p_{S,\sigma^2}(\varphi|y). \end{aligned} \tag{5.6}$$

Then we have:

1. Equation (5.6) is the signal model in phase data space, which is a convenient representation in physics.
2. The conditional density

   $$p_{S,\sigma^2}(\varphi|y) = \frac{1}{2\pi I_0(Sy/\sigma^2)}\exp\Big(\frac{Sy}{\sigma^2}\cos(\varphi)\Big), \quad \varphi \in [0, 2\pi), \tag{5.7}$$

   is an instance of the symmetric von Mises distribution on the circle, see Fisher et al. (1987) [Chapter 4.3.2]. Note also that for $y = 0$ we obtain the Gaussian likelihood

   $$p_{S,\sigma^2}(\varepsilon_r = -S, \varepsilon_i = 0) = \frac{y}{2\pi\sigma^2}\exp\Big(-\frac{S^2}{2\sigma^2}\Big),$$

   and in such a case the augmentation is not needed.
3. When the MR signal is $S = \exp(Z\boldsymbol{\theta})$,

   $$L(\beta|y,\varphi,Z) = \frac{y}{2\pi\sigma^2}\exp\Big(-\frac{1}{2\sigma^2}\big(y^2 + \exp(2Z\boldsymbol{\theta}) - 2\exp(Z\boldsymbol{\theta})y\cos(\varphi)\big)\Big), \tag{5.8}$$

   which is the complete-data likelihood of a GLM with Gaussian response up to a constant depending on the observation and log link function

   $$g(\mathbb{E}(Y,\varphi|\beta,Z)) = \log(\mu) = Z\boldsymbol{\theta},$$

   conditionally on $\cos\varphi$, and the mean of the response, $\mu = \exp(Z\boldsymbol{\theta})$.

5.2 The EM algorithm for fast estimation

The EM algorithm (Sundberg , 1974, Dempster et al. , 1977, Robert and Casella , 2004) is usually considered as an efficient alternative to overcome difficulties in maximizing the likelihood, especially concerning latent variables as in mixture modeling. It is particularly straightforward when applied to models belonging to the exponential family (Sundberg , 1974). This thesis elaborates the original implementation of the EM algorithm in DTI in connection with data augmentation.

The general idea of the EM algorithm under missing data models can be summarized as follows. Consider a statistical model $\big(p_\vartheta(y), \vartheta \in \beta\big)$, where $\beta \subseteq \mathbb{R}^d$, and the likelihood of the observed data $y = (y_1, \ldots, y_n)$ is expressed as the marginal of the joint likelihood

$$p_\vartheta(y) = \int_{\mathcal{Z}} p_\vartheta(z, y)\, dz.$$

Here $z = (z_1, \ldots, z_n) \in \mathcal{Z}$ and z_i are interpreted as latent variables. When $\mathcal{Z}$ is discrete, we replace integrals by sums. In the EM algorithm, starting with an initial value $\vartheta^{(0)} \in \beta$, we calculate the expectation of the log-likelihood function with respect to the conditional distribution of z given y under the current estimate of the parameter ϑ at step k

$$\mathbb{E}_{\vartheta^{(k)}}\big(\log p_\vartheta(z, y)\big|y\big), \tag{5.9}$$

being the expectation step. In the maximization step, we compute

$$\vartheta^{(k+1)} = \arg\max_{\vartheta\in\beta}\Big\{\mathbb{E}_{\vartheta^{(k)}}\big(\log p_\vartheta(z, y)\big|y\big)\Big\} = \arg\max_{\vartheta\in\beta}\Big\{\int_{\mathcal{Z}} \log p_\vartheta(z, y) p_{\vartheta^{(k)}}(z|y)dz\Big\} \tag{5.10}$$

(Dempster et al. , 1977), where the integration is with respect to the conditional density

$$p_{\vartheta^{(k)}}(z|y) = \frac{p_{\vartheta^{(k)}}(z, y)}{p_{\vartheta^{(k)}}(y)}.$$

By Jensen's inequality, the Kullback relative entropy of the conditional distribution $p_\vartheta(z|y)$ related to $p_{\vartheta^{(k)}}(z|y)$, given by

$$K(\vartheta^{(k)}, \vartheta|y) := \mathbb{E}_{\vartheta^{(k)}}\left(\log\left(\frac{p_{\vartheta^{(k)}}(z|y)}{p_\vartheta(z|y)}\right)\middle|y\right) = \int_{\mathcal{Z}} \log\left(\frac{p_{\vartheta^{(k)}}(z|y)}{p_\vartheta(z|y)}\right) p_{\vartheta^{(k)}}(z|y)dz\ ,$$

is non-negative, which implies

$$\begin{aligned}&\log p_\vartheta(y) - \log p_{\vartheta^{(k)}}(y) \geq \\ &\int_{\mathcal{Z}} \log\big(p_\vartheta(z, y)\big) p_{\vartheta^{(k)}}(z|y)dz - \int_{\mathcal{Z}} \log\big(p_{\vartheta^{(k)}}(z, y)\big) p_{\vartheta^{(k)}}(z|y)dz\ ,\end{aligned} \tag{5.11}$$

and consequently

$$\log p_{\vartheta^{(k+1)}}(y) \geq \log p_{\vartheta^{(k)}}(y) \ .$$

In other words, the EM step do not decrease the marginal likelihood of y. It follows also from Equation (5.11) that fixing a ϑ-subvector and maximizing with respect to the remaining ϑ-coordinates do not decrease the marginal likelihood of y. The EM algorithm is iterated until convergence to a fixed point $\vartheta^{(\infty)}$, a local maximum of the marginal likelihood $p_\vartheta(y)$. When the local maximum is the global one, $\widehat{\vartheta}_{ML} = \vartheta^{(\infty)}$ is the maximum likelihood estimator of the parameter. In fact, it turns out that the limiting results from the EM coincide with the ML estimates. The advantage of the EM algorithm is that, for some smart choices of the augmented data z and the joint density $p_\vartheta(z, y)$, the maximization step in Equation (5.10) can be calculated more easily than maximizing the marginal likelihood $p_\vartheta(y)$ directly, especially in cases where the latter is hard to evaluate. In cases where $p_\vartheta(z, y)$ can be maximized using a standard software, the algorithm is easy to implement. A minor drawback is that a large number of iterations are needed.

5.3 EM in diffusion MRI

We illustrate the advantages of the EM algorithm in DTI by the Poisson data augmentation. In details, the data augmentation is in accordance with the likelihood of the signal measurement in Equation (5.5): Let $t = S_0^2 \exp(2Z\boldsymbol{\theta})/2\sigma^2$, the complete data log-likelihood is then expressed as

$$Q := \log\big(p_{t,\sigma^2}(N = n, y)\big) = c(y, n) + n\log(t) - (n+1)\log(\sigma^2) - t - \frac{y^2}{2\sigma^2}, \quad (5.12)$$

where $c(y, n) = n\log(y^2) - 2\log(n!) - (n+1)\log(2)$ does not depend on (t, σ^2) and will be omitted in the M step. The EM algorithm proceeds in two steps when maximizing the likelihood: in the E step, given the current parameter estimates $(\boldsymbol{\theta}^{(k)}, S_0^{2(k)}, \sigma^{2(k)})$, we update the conditional expectation of the augmented data by

$$\mathbb{E}_{t^{(k)},\sigma^{2(k)}}\big(N\big|y\big) = \frac{\tau^{(k)}\ I_1\big(2\tau^{(k)}\big)}{I_0\big(2\tau^{(k)}\big)} \qquad \text{with} \qquad \tau^{(k)} = y\sqrt{\frac{t}{2\sigma^{2(k)}}}.$$

In the M step, σ^2 and S_0^2 are updated by the recursions by the modes from their marginals. In the E step, we can update the tensor parameter $\boldsymbol{\theta}$ by applying the Fisher scoring method, for more details see **PII**. Moreover, the EM algorithm also works in the Bayesian framework to find point estimates, and we can invoke the maximum a posteriori (MAP), more discussion can be found in **PII**.

6 BAYESIAN MODELING, COMPUTATION AND REGULARIZATION

Bayesian theory stems from the Bayes formula

$$p(\beta|y) = \frac{p(\beta, y)}{p(y)} = \frac{p(\beta)p(y|\beta)}{p(y)}, \tag{6.1}$$

where β stands for the vector of unknown quantities and y for data. When using this formula, the right hand side needs to be specified despite that the normalizing constant $p(y) = \int p(\beta)p(y|\beta)d\beta$ may remain unknown. Therefore, in most cases we use the non-normalized form

$$p(\beta|y) \propto p(\beta)p(y|\beta), \tag{6.2}$$

see e.g. Gelman et al. (2014), Robert and Casella (2004), Bernardo and Smith (1994). Additionally, the Bayes formula implies the derivation of two posterior densities expressed in Equation (5.2) and Equation (5.7) from the previous chapter. In particular, in terms of the reinforced Poisson distribution it represents the conditional distribution in Equation (5.2). It is defined by virtue of the fact that, after normalization any convergent series with positive terms becomes a probability distribution, and in this case we know the normalizing constant.

Compared with the frequentist statistical inference, such as the maximum likelihood method, the key difference to the Bayesian approach is that in the latter the inference on the parameters is based on the posterior density $p(\beta|y)$, depending not only on the likelihood but also on the prior and assumptions on the unknown parameter encoded by the prior distribution $p(\beta)$.

6.1 Prior selection

Specification of the prior for the unknown parameters is necessary in building a Bayesian model. In what follows we are going to consider three classes of priors in Bayesian modeling, which are used in this thesis, see **PI**.

Conjugate priors. When the prior knowledge is allowed, the primary choice of the prior distribution is from a conjugate family[1]. This is because the conjugate priors lead to computational tractability, achieving the property that the prior is chosen such that both the prior and the posterior belong to the same family of distributions. This choice depends on the likelihood. This principle stems from the Bayes formula in Equation (6.2) that conjugate priors give a closed-form posterior, avoiding the tedious computation of the normalizing constant. However, conjugate priors exist only when the likelihood belongs to an exponential family, see e.g. Robert and Casella (2004) [Chapter 1.6], Schervish (1995) [Chapter 2].

Let the signal intensity be $S = S_0 \exp(Z\boldsymbol{\theta})$, so that we can distinguish the non-diffusion weighting signal S_0 separately from the tensor $\boldsymbol{\theta}$. Recall the Poisson DA model in count data space: the joint likelihood in Equation (5.5), and consider S_0 as the parameter of interest, while the other parameters are assumed to be fixed. Then the likelihood is reduced to the Poisson part

$$\mathcal{L}(S_0^2|\boldsymbol{\theta}, \sigma^2, N_i, Z_i,\ i = 1, \cdots, m) = const \times (S_0^2)^a \exp(-bS_0^2), \tag{6.3}$$

with

$$a = \sum_{i=1}^{m} N_i, \qquad b = \frac{1}{2\sigma^2} \sum_{i=1}^{m} \exp(2Z_i\boldsymbol{\theta}).$$

When $p(y_i, N_i,\ i = 1, \cdots, m|S_0^2) = \mathcal{L}(S_0^2|\boldsymbol{\theta}, \sigma^2, N_i, Z_i,\ i = 1, \cdots, m)$ is viewed as a function of S_0, which is the Gamma density with hyperparameter $\zeta = (a, b)$, belonging to the exponential family. We can choose the prior $p(S_0^2) \sim \mathrm{G}(c_1, c_2)$, where G stands for the gamma distribution. Here c_1 and c_2 denote the shape and inverse scale parameters, respectively. By the conjugate property, the full conditional posterior of S_0^2 given the other parameters is

$$p(S_0^2|\boldsymbol{\theta}, \sigma^2, N_i, Z_i,\ i = 1, \cdots, m) \sim \mathrm{G}(a + c_1, b + c_2), \tag{6.4}$$

again a Gamma distribution, and we know the normalizing constant analytically.

In the case the prior information on S_0^2 is weak, one can choose the hyperparameters c_1 and c_2 to be small positive constants, leading to a relatively flat prior distribution over a large range. Then the posterior distribution is dominated by the likelihood. On the other hand, in order to model a large dataset more accurately, it is natural to involve multiple parameters and build a hierarchical model so that we have enough parameters to explain the variation. In doing that, we can assign prior distributions also to the hyperparameters.

---

[1] Conjugate families can be denoted by $\mathcal{P} = \{\beta \mapsto \pi(\beta|\zeta) : \zeta \in \mathcal{E}\}$, where $\mathcal{E}$ denotes a Euclidean space, and ζ is a hyperparameter vector, including such as shape and scale parameters, or mean, for instance. The conjugate property is if $p(\beta) = f(\beta|\zeta_0) \in \mathcal{P}$ for some ζ_0, then the posterior will be formed by $p(\beta|y) = f(\beta|\zeta_1)$, and $\zeta_1 \in \mathcal{E}$ depends on the response y, see Koistinen (2010) for a detailed discussion or Bernardo and Smith (1994) for a rigorous definition.

Non-informative priors A key feature in the Bayesian learning process is that it combines the prior and the likelihood. Therefore, the specification of the prior plays a crucial role in modeling, representing the prior beliefs about the underlying problem. Non-informative priors attempt to avoid subjective elicitation (information) about the parameters of interest into the model, see e.g. Robert and Casella (2004), Box and Tiao (1992). Further, improper priors form a non-informative class, commonly appearing in Bayesian modeling, with infinite mass $\int p(\beta)d\beta = +\infty$. Note that the prior is allowed to be improper but the posterior should always be a proper probability distribution.

In DTI we set σ^2 to have a scale-invariant improper prior with density $p(\sigma^2) \propto 1/\sigma^2$. By Bayes theorem, the full conditional posterior density of σ^2, given $\boldsymbol{\theta}$, S_0 and the augmented data N_i is

$$p(\sigma^2|\theta, S_0, N_i, y_i, Z_i,\ i = 1, \cdots, m) \propto$$
$$\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\{y_i^2 + \exp(2Z_i\boldsymbol{\theta})S_0^2\}\right)(\sigma^2)^{-\left(1+\sum_{i=1}^{m}(2N_i+1)\right)}, \tag{6.5}$$

which is an inverse Gamma distribution with shape and rate parameters

$$\sum_{i=1}^{m}(2N_i + 1) \qquad \text{and} \qquad \frac{1}{2}\sum_{i=1}^{m}\big(y_i^2 + \exp(2Z_i\boldsymbol{\theta})\big), \ \text{respectively.}$$

Furthermore, the improper prior $p(\sigma^2) \propto 1/\sigma^2$ can also be considered as a conjugate prior of the inverse Gamma form in the extreme case that the values of the shape and inverse scale parameters tend to zero. By virtue of the conjugate properties through the joint likelihood in Equation (5.5) viewed as a function of σ^2, we get the inverse Gamma as the full conditional posterior distribution of σ^2 that belongs to the same family distributions as the prior. It should be noticed that an improper prior may lead to a non-integrable (joint) posterior distribution.

Informative priors An informative prior transmits definite, or at least reasonable information to the model. Basser and Pajevic (2003) suggested a multivariate normal density for tensor matrix $\overline{D}$, which preserves the algebraic relations and geometric structure of the tensor elements. In their work, the distribution of the tensor matrix $\overline{D}$ was addressed to be isotropic *iff* it has a density of the form

$$\pi(D) = \frac{\eta^{5/2}\sqrt{\eta + 3\lambda}}{(\pi\sqrt{2})^3}\exp\left(-\frac{1}{2}\left(\eta\text{tr}(D^2) + \lambda\{\text{tr}(D)\}^2\right)\right), \tag{6.6}$$

where the hyperparameters η and λ should follow the constraints $\eta > 0$ and $\lambda > -\eta/3$, and "tr" stands for the matrix trace. On the basis of this suggestion, an isotropic centered Gaussian prior $\mathcal{N}(0, \Omega^{-1})$ can be specified for the tensor

parameter $\boldsymbol{\theta} \in \mathbb{R}^6$, where the precision matrix

$$\Omega = \begin{pmatrix} \lambda+\eta & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+\eta & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+\eta & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\eta & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\eta & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\eta \end{pmatrix}. \tag{6.7}$$

is controlled by the hyperparameters η and λ.

From the statistical perspective, the specification of the prior being normally distributed is widely used because of its important properties, see Fukunaga (2013) for more details. Moreover, the Gaussian prior obtains maximum entropy (see e.g. Friedman et al. (2001)) among all distributions with support $\mathbb{R}^d$ and with given covariance matrix Ω: It minimizes the amount of prior information transmitted to the posterior. The prior can be easily extended for the tensor with any higher order ($d \geq 6$), see an example with the case of 4th order tensor in **PI**.

6.2 Markov chain Monte Carlo sampling

Bayesian modeling renders insight into how to interpret parameters of interest in a regression problem via either point estimates or the probability distributions of the parameters, which is achieved by simulation schemes. We shall start with the latter one based on the fully Bayesian approach and discuss related sampling algorithms.

Markov chain Monte Carlo (McMC) is a class of dynamic algorithms for simulating samples from a target distribution. It involves three basic set-up concepts:

1. Let us first define the transition (probability) kernel of a homogeneous Markov chain $\{\Theta^{(t)}\}$ on the state space $\mathbb{S}$ as $K(\beta, d\alpha) = P(\Theta^{(t+1)} \in d\alpha | \Theta^{(t)} = \beta)$ for $\alpha, \beta \in \mathbb{S}$. The transition kernel is said to be reversible w.r.t. the probability distribution π if the Markov chain satisfies the local detailed balance condition

$$\pi(d\alpha)K(\alpha, d\beta) = \pi(d\beta)K(\beta, d\alpha), \qquad \alpha, \beta \in \mathbb{S}. \tag{6.8}$$

   This implies that π is an invariant (stationary) distribution and the global balance condition $\int\limits_{\alpha \in \mathbb{S}} \pi(d\alpha)K(\alpha, d\beta) = \pi(d\beta)$ holds (Chung , 1967, Häggström , 2002, Banerjee et al. , 2004). Note that the local balance implies the global balance, but the opposite implication does not hold.
2. If the kernel K is $\pi-$irreducible and aperiodic with the invariant distribution π, then for π-almost all β,

$$\lim_{t \to \infty} K^t(\beta, d\alpha) = \pi(d\alpha) \text{ in total variation,}$$

see Tierney (1994) [Theorem 1]. By Nummelin (1984) [Proposition 6.3], the chain is ergodic. About irreducibility and aperiodicity, rigorous mathematical definitions can be found e.g. in Tierney (1992), Nummelin (2002), Roberts and Rosenthal (2004). Moreover, the ergodic theorem ensures that the sample path average converges to the expectation under the distribution π when it is finite,

$$1/T \sum_{t=0}^{T} f(\Theta^{(t)}) \to \mathbb{E}_{\pi}(f), \text{ as } T \to \infty, \tag{6.9}$$

see Tierney (1994) [Theorem 3]. It should be noticed that although the law of large numbers (LLN, see e.g. Durrett (2010)) follows from the ergodic theorem, here we do not need the sequences $f(\Theta^{(t)})$ to be independent.

3. Working with an McMC algorithm involves several issues: 1) the choice of the kernel K; 2) the length of the burn-in period, which brings the Markov chain close to the equilibrium after starting from an arbitrary state, and should be ignored when computing the empirical average; 3) the practical running time T: if the choice of T is not large enough, the chain may not mix well enough, whilst if T is too large, then the computational burden will be considerable; 4) the simulation variance that controls the accuracy of the estimates of expectations obtained through McMC. This is why we commonly see variance reduction techniques also to be embarked in some McMC strategies. Adaptive McMC is among such techniques.

In summary, McMC may be interpreted as that the transition kernel K^t is ergodic with limiting distribution π, for any initial stage $\Theta^{(0)}$ of the chain, the empirical mean expressed in Equation (6.9) converges to the expectation w.r.t. the equilibrium distribution π. This is achieved by a long and realizable chain (Besag et al. , 1995).

6.3 Gibbs sampler

Two popular McMC methods involved in this thesis are Gibbs sampler and the Metropolis-Hastings algorithm. Gibbs sampler (Geman and Geman , 1993) is the best known among the McMC algorithms. The idea behind Gibbs sampling is that at each recursive cycle, we draw a sample β_j from its full conditional distribution given the other components $\beta_{-j} := \{\beta_1, \cdots, \beta_{j-1}, \beta_{j+1}, \cdots, \beta_n\}$, where n is the number of components of β. For instance, in DTI we draw $(S_0^2)^{t+1}$ and $(\sigma^2)^{t+1}$ at state t from their full conditional posteriors. The reader should be aware that Gibbs sampling may contain a dangerous situation that "Gibbs sampler will lead to seemingly reasonable inferences about a nonexistent posterior distribution". Hobert and Casella (1996) argued and pointed out that this has appeared in some published works. This danger appears mainly when the prior is improper, in which situation a mathematical demonstration of the propriety of the

(joint) posterior is necessary. An example regarding to the prior distribution of the tensor can be found in **PI**.

6.4 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm (Metropolis et al. , 1953, Hastings , 1970) is an alternative way of setting up an McMC method, which is a universal tool in statistical inference for exploring high-dimensional probability distributions. The idea behind MH is to generate a Markov chain by an acceptance/rejection rule, converging to the target distribution.

Let $\pi(\Theta) = c^{-1}p(\Theta|y)$ be the target probability density for a parameter $\Theta \in \mathbb{R}^d$, where the normalizing constant

$$c = \int\limits_{\mathbb{R}^d} p(\Theta)p(y|\Theta)d\Theta < \infty$$

may be unknown. Starting from state (t), we draw a proposal Θ' from a user-defined proposal density $q(\Theta^{(t)}|\Theta')$ as a suggested value in the successive state of a Markov chain in the parameter space, and calculate the MH acceptance ratio

$$r(\Theta^{(t)}, \Theta') := \min\left\{\frac{p(\Theta'|y)q(\Theta^{(t)}|\Theta')}{p(\Theta^{(t)}|y)q(\Theta'|\Theta^{(t)})}, 1\right\}. \tag{6.10}$$

With probability $r(\Theta^{(t)}, \Theta')$ the proposed value is accepted and set to be $\Theta^{(t+1)} = \Theta'$; otherwise we keep the old value $\Theta^{(t+1)} = \Theta^{(t)}$.

It is straightforward to check that the resulting density of the transition kernel

$$k(\Theta^{(t)}, \Theta') = r(\Theta^{(t)}, \Theta')q(\Theta'|\Theta^{(t)}) + (1 - A(\Theta^{(t)}))\delta_{\Theta^{(t)}}(\Theta'),$$

with $\delta_{\Theta^{(t)}}$ denoting the point mass at $\Theta^{(t)}$ and $A = \int r(\Theta^{(t)}, \Theta')q(\Theta'|\Theta^{(t)})d\Theta'$, satisfies the detailed balance condition

$$\pi(\Theta^{(t)})r(\Theta^{(t)}, \Theta')q(\Theta'|\Theta^{(t)}) = \pi(\Theta')r(\Theta', \Theta^{(t)})q(\Theta^{(t)}|\Theta'),$$

and that the Markov chain (Θ^t) generated by MH is reversible w.r.t. the target distribution $\pi(\beta)$, see Robert and Casella (2004) [Chapter 7], Tierney (1992), Nummelin (2002).

One main issue in the MH algorithm is the choice of the proposal density. Namely, the construction of McMC proposals with good mixing properties, especially in high-dimensional cases, crucially determines the success of the McMC procedure. Robert and Casella (2004) provide further discussion about this issue. In DTI, we suggested a Gaussian proposal

$$q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \big(\Omega + I(\widehat{\boldsymbol{\theta}})\big)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right) \tag{6.11}$$

for updating the tensor parameter $\boldsymbol{\theta}$. This follows the rather convenient scheme in Bayesian framework, where the posterior distribution of the tensor is assumed to be unimodal and symmetric.

To summarize, Gibbs sampler has a natural appeal due to its easier form than the MH algorithm and can be less efficient than the MH in computation. What is most important, the Gibbs algorithm needs the full conditional posteriors from which we can directly draw samples, whilst the MH needs only the unnormalized posterior. MH is a rather universal method and allows a fairly wide range of choices for the instrumental distribution (the proposal), especially when the full conditionals are unknown. However, as we indicated, to find an ideal proposal for getting good mixing properties is sometimes challenging. Furthermore, Gibbs sampler can be viewed as a special case of the MH algorithm, setting the proposal to be equivalent with the full conditionals, in which case the acceptance rate will be equal to 1. In addition, both algorithms can be combined together as a hybrid strategy reaching the same limiting (equilibrium) distribution to solve complicated problems as described in our example.

6.5 Adaptive McMC

As was previously mentioned, it is not an easy task to find a good proposal for a particular problem concerning complex and high-dimensional problems. A common statistical recipe is to use the random walk Metropolis (RWM) algorithm, due to its easy and efficient implementation. However, a multivariate normal proposal, $q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \sim \mathcal{N}(\boldsymbol{\theta}^{(t)}, \Sigma^{(t)})$ with $\boldsymbol{\theta} \in \mathbb{R}^d$, as we illustrated, does not belong to the RWM. It may occur that if the variation allowed by the proposal is too small, the convergence rate is very slow and the chain does not achieve the stationary state in a reasonable time. Instead, if the variation in the proposal is too large, then the amount of rejection is untenable. In both cases the inference on the simulated chain is unreliable, see e.g. by Atchadé et al. (2009). Gelman et al. (1996, 2014) suggested to use adaptive algorithms, where the proposal distribution is allowed to depend on the whole past of the process. For example, we may choose the covariance matrix in the Gaussian proposal as $\Sigma^{(t)} = (2.38^2)/d \cdot \widehat{\Sigma}^{(t)}$, where $\widehat{\Sigma}^{(t)}$ is the empirical covariance matrix. This choice is proved to be optimal for some toy models in Roberts et al. (1997), Roberts and Rosenthal (2001).

6.6 Variational Bayes approximation

Variational Bayes (VB) is an approximative method for posterior computation. Unlike in McMC, approximating the target distribution by the empirical distribution of a Markov chain exploring the state space that typically needs long iterations, VB on the other hand intends to solve an optimization problem leading to

marginal posterior approximations. It is often preferred as an efficient algorithm employed in high-dimensional regression problems in connection with big data, for instance as a tractable signal processing algorithm to infer on parameters. The idea of the mean-field variational Bayes framework (VB) is to approximate the joint posterior distribution of the parameter $\beta = (\beta_1, \ldots, \beta_m)$,

$$p(\beta|y) = p(\beta)p(y|\beta)/p(y),$$

simply by a product of probability distributions $\widehat{q}(\beta) = \widehat{q}_1(\beta_1)\widehat{q}_2(\beta_2)\cdots\widehat{q}_m(\beta_m)$. Note that this factorized solution usually appears in the EM based VB algorithm (see e.g. Beal et al. (2002)) to make the variational optimization problem easy to solve, but it is not a general restriction in the VB algorithm. The fixed-form variational Bayes (Saul and Jordan , 1996, Salimans and Knowles , 2013) as an alternative choice can be adopted in a wide variety of models without such a factorization.

The VB marginals $q(\beta)$ in the approximation of $p(\beta|y)$ are achieved minimizing the Kullback divergence (KL)

$$\mathrm{K}\big(p(\cdot|y) \parallel q(\cdot)\big) = \int q(\beta) \log\left(\frac{q(\beta)}{\pi(\beta|y)}\right) d\beta, \tag{6.12}$$

see Kullback and Leibler (1951). The minimum is obtained by solving iteratively the VB-marginal recursions

$$\widehat{q}_j^{(t+1)}(\beta_j) \propto \exp\left(\int \log p(y,\beta) \prod_{i\in\{-j\}} \widehat{q}_i^{(t)}(\beta_i) d\beta_1 \cdots d\beta_{j-1}\, d\beta_{j+1} \cdots d\beta_m\right), \tag{6.13}$$

as in Šmídl and Quinn (2006). This equation states that the essence of VB is to approximate the marginal posterior $p(\beta_j|y)$ by the VB marginal $\widehat{q}_j(\beta_j)$. The method therefore sacrifices estimation accuracy through the (usually) facilitative approximative form of the marginals, the VB marginal, to gain in computational efficiency. The detailed explanation and derivation of the VB updating algorithm can be found in a wide choice of literature e.g. Ormerod and Wand (2010), Jaakkola and Jordan (2000).

Prerequisites for VB Equation (6.13) can be further written as

$$\widehat{q}_j(\beta_j) \propto \exp\left(\mathbb{E}_{\beta_{-j}} \log p(\beta_j|y,\beta_{-j})\right).$$

Apparently, the right hand side in this expression is in connection with Gibbs sampler through the full conditionals $p(\beta_j|y,\beta_{-j})$. The vital prerequisite for which the algorithm can work is the tractability of the VB marginals. In other words, we are able to compute the moments or expectations of VB parameters β_{-j} that appear in $\widehat{q}_j(\beta_j)$. Secondly, the full conditionals $p(\beta_j|y,\beta_{-j})$ should be either tractable or one should be capable of dealing with those. In the latter case, the data augmentation and/or the Laplace approximation may help in configuring

the intractable full conditionals, see our example in the next paragraph. Thirdly, it is worthwhile to point out that improper priors may lead the VB approximation to have an unwanted performance, which may yield unreliable inference or a convergence failure, see Zhao (2013) [Chapter 3] for concrete examples. Therefore, it is necessary to check the propriety of the posterior before employing the algorithm.

Stopping criteria Without a proper tolerance, the VB algorithm can be even less efficient than the McMC. The stopping criteria for the VB algorithm is based on the left hand side in the information inequality

$$\int \log\left(\frac{p(\beta, y)}{\widehat{q}^{(t)}(\beta)}\right)\widehat{q}^{(t)}(\beta)d\beta \leq \log p(y)$$

with the joint distribution of the parameters, the response $p(\beta, y) = \pi(\beta)p(y|\beta)$, and the marginal density of the response $p(y)$. The left hand side should be increasing between consecutive iterations, and when it does not increase anymore up to numerical tolerance, the algorithm can be stopped. In **PIV** we proposed a VB algorithm with DKI in diffusion MRI.

6.7 Bayesian regularization and GMRF

A salient goal in this context is to restore the diffusion tensor derived images from the original degraded diffusion MRI data by removing the oscillation caused by artefacts from structural noise such as bulk movement and random noise, and to preserve important structural information of the images. The Bayesian regularization as one of denoising tools in image studies is easy to apply in Bayesian framework with any model of signal decay where the data have the Rician distribution.

Bayesian regularization as a smoothing technique, in the spirit suggested by Geman and Geman (1993), has been successfully applied in image restoration, see e.g. Frandsen et al. (2007), Krissian and Aja-Fernández (2009). In this approach the most commonly used prior model that describes the relevant image attributes, e.g. interdependencies between associated pixels in a given neighborhood system, is the Markov random field (MFR), or the so-called Gibbs-type distribution [2]. The accurate definition and interpretation of MRF can be found e.g. in Hammersley and Clifford (1971), Cross and Jain (1983), where the authors formulate three postulates: positivity, Markov property through a neighborhood condition, consistency condition and homogeneity. The principal property of MRF is conditional independence. The notations stem from the graph

[2] A Gibbs distribution may be represented by a probability function $\pi(w) = \frac{1}{T_1}\exp(-U(w)/T_2)$, where T_1 and T_2 are two normalizing constants, $U(\cdot)$ refers to the energy function, see Hammersley and Clifford (1971), Geman and Geman (1993), Geman and Graffigne (1986), Frandsen et al. (2007) for a detailed explanation and specific examples.

theory, see e.g. Golumbic (2004). For any set W, we define the exterior boundary of W by $\partial W := \{w \in V \setminus W : \exists v \in W \text{ with } w \sim v\}$, where V is the set of all the pixels (also called *lattices* in mathematics, *nodes or vertices* in graph theory) in a given image, and $W \subseteq V$. The closed neighborhood [3] $\overline{W}$ is defined as $\overline{W} := W \cup \partial W$. Additionally, we use $\partial\{v\} := \{w \in V : w \sim v\}$ to indicate the neighborhood of v. Here "$\sim$" stands for a neighborhood relation, $V \setminus W$ denotes a set $\{v \in V : v \notin W\}$. Local and global Markov properties given by

$$\boldsymbol{\theta}_W \perp\!\!\!\perp \boldsymbol{\theta}_{V\setminus\overline{W}} \mid \boldsymbol{\theta}_{\partial(W)}$$

are in fact equivalent according to the Hammersley-Clifford terminology (Hammersley and Clifford , 1971, Besag , 1974, Clifford , 1990).

Gaussian Markov random fields (GMRF) The Gaussian assumption is extensively used in statistical models and is the most common choice in a MRF due to its convenient computational and tractable properties. When the tensor parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n) \in \mathbb{R}^n$ has a multivariate normal distribution with mean μ and precision matrix (inverse covariance matrix) $\mathbf{Q}$, then a GMRF w.r.t a finite undirected graph $G = (V, \mathcal{E})$ is defined by

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-n/2} det(\mathbf{Q})^{1/2} \exp(-1/2(\boldsymbol{\theta} - \mu)^T \mathbf{Q}(\boldsymbol{\theta} - \mu)^T),$$

see also Rue and Held (2005) [Chapter 2], where V is the set of all voxel of interest, $\mathcal{E}$ denotes the set of edges $\{(v, w) \in \mathcal{E}, \text{ with } v, w \in V, v \neq w\}$, and det is the determinant matrix operator. Moreover, the precision matrix $\mathbf{Q} > 0$ and $\mathbf{Q}_{vw} \neq 0 \Leftrightarrow \{v, w\} \in \mathcal{E}$ for all $v \neq w$, and in DTI $n = 6$.

6.8 Nearest neighboring system in 3D neural networks

In a true scenario, neighboring pixels usually form an ensemble of similar intensity. The most common neighborhood structure is the nearest neighboring, see e.g. Hans et al. (2007). In this thesis, we count the nearest six pixels as neighbors of v and in the three-dimensional (3D) spatial space, depicting in Figure 16. The triangle nearest neighbors therefore do not fall into our neighborhood configurations, see more details in **PI**. Additionally, Figure 16 illustrates a simple graph where the set of cliques [4] consists of all adjacent pairs of edges.

In our neighboring system, we apply the neighborhood structures to blocks of pixels, i.e., a cubic lattice contains a set of pixels, and we restrict the blocks always to be rectangular. In Bayesian analysis, we can set a tuning parameter to control the size of the block, which is within the range from one single pixel to the

[3] The closed neighborhood is that for any lattice $v \in W$, v *per se* is included in the neighborhood system.

[4] A clique is the building block to configure a MRF, its elaboration can be found in e.g. Hammersley and Clifford (1971), Geman and Geman (1993), Clifford (1990).

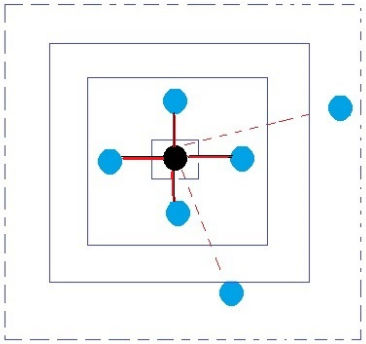


FIGURE 16 Neighborhood structure of one pixel. The black circle indicates the pixel v, the blue color points are the neighbors of v. The black-red lines show the edges in this simple graph, where the dashed lines connect the neighboring pixels from the front and the back in the 3D network.

cubic lattice containing all the pixels from a given image. However, the reader should bear in mind that the realistic neighborhoods must be small enough to ensure feasible computational loads and big enough to reach the goal of image restoration (Cross and Jain , 1983, Geman and Geman , 1993).

6.9 GMRF for DT

In diffusion tensor profile, these voxels are mutually connected and locate on several consecutive slices (layers) of the brain. Water molecules diffuse along the underlying fibres across several voxels and, as a consequence, the tensors are not independent. The correlation between two tensors / blocks of tensors depends not only on the distance but also on the location. This is why we need neighboring configurations to detect the tensors from the MR images which are corrupted and/or contain missing observations. Figure 17 illustrates tensor fields of 2nd order and 4th order, respectively, from a region of interest (ROI) without (left) and with (right) regularization. Both pairs of figures show the regularization effects, which are more obvious in the 4th order tensor field than in the 2nd order one.

An isotropic Gaussian pairwise difference prior for DT The neighboring system can be simply interpreted by a homogeneous (stationary and isotropic) Gaussian Markov random fields (GMRF), see Banerjee et al. (2004). Under the basis of the formalization of the isotropic Gaussian prior in Equation (6.6), we propose a proper isotropic prior for a GMRF of 3×3 symmetric matrices $(D(v) : v \in V)$

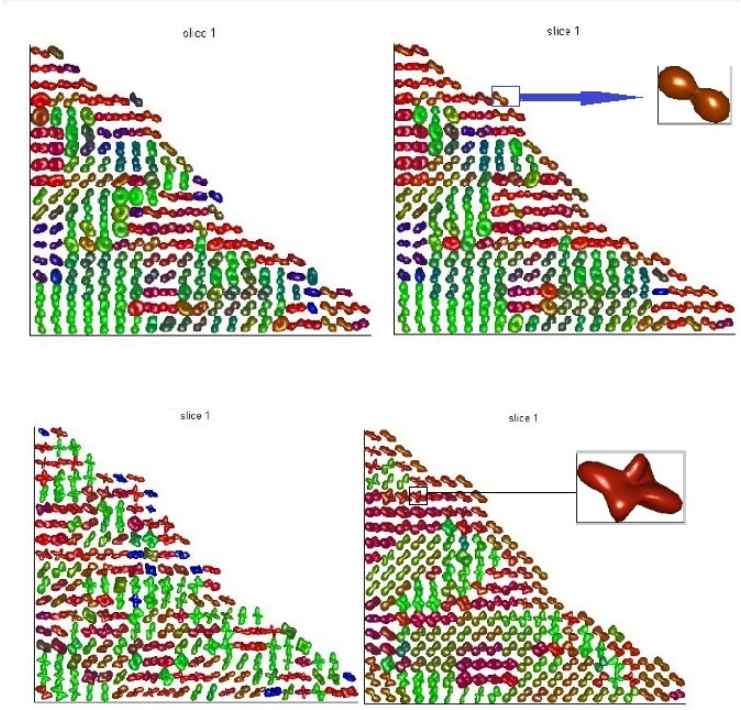


FIGURE 17 The upper two figures describe the 2nd order tensor fields (with zoom) of a ROI from a slice of a healthy human brain without (left image) and with (right one) regularization. The bottom two figures are the 4th order tensor fields from the same ROI of 4th order. The color-coded representation shows the main direction of the principal eigenvalue of the tensor: red, left-right; green, anterior-posterior; blue, superior-inferior.

for the 2nd order tensor fields in the 3D spatial space, which is given by

$$\pi(D(v) : v \in V) =$$
$$(2\pi)^{-|V|d/2} \det(\Omega)^{|V|/2} \det(I_V + \rho L_V)^{d/2} \exp\left(-\frac{1}{2}\theta^\top \{(I_V + \rho L_V) \otimes \Omega\}\theta\right) \quad (6.14)$$
$$\propto \exp\left(-\frac{1}{2}\sum_{v\in V}\left\{\eta \text{Trace}(D(v)^2) + \lambda\{\text{Trace}(D(v))\}^2\right\}\right.$$
$$\left. - \frac{\rho}{2}\sum_{v\sim w}\left(\eta \text{Trace}(\{D(v) - D(w)\}^2) + \lambda\{\text{Trace}(D(v) - D(w))\}^2\right)\right), \quad (6.15)$$

where $\rho \geq 0$ tunes the dependence between tensors at different voxels, L_V denotes the Laplacian matrix of the graph V, and $|V|$ counts the number of nodes in the set V. We can omit the normalizing constant

$$(2\pi)^{|V|d/2} det(\Omega)^{|V|/2} \det(I_V + \rho L_V)^{d/2}$$

in the calculation. An analogous construction can be found in Kaipio and Somersalo (2006) termed as the least-squares Tikhonov regularization for penalizing the likelihood. Bayesian regularization has been applied in **PI** and **PIV**, where we have more details.

PART III

7 CONCLUSION AND DISCUSSION

7.1 Two schemes of DA

In this thesis, we propose two different schemes of data augmentation on the same problem. It is a natural concern for the distinction of the two strategies with detailed statistical interpretation, though they are intended to solve the same problem.

1. Firstly, the two DA schemes operate in the different data spaces.
2. DA in the count data space reduces the nonlinear regression problem *partially* into the GLM framework, meaning that the complete-data likelihood expressed in Equation (5.5) is not the rigorous one of GLMs: When the tensor $\boldsymbol{\theta}$ is considered as the only unknown parameter, the likelihood then falls into the GLM framework with Poisson response; whereas in the phase data case, the Rician likelihood has completely transferred into a GLM framework with the complete-likelihood in Equation (5.8) by augmenting the phase data φ.
3. We have derived the conditional probability distributions of both augmented data, expressed in Equation (5.2) and Equation (5.7). The relevance of these two schemes depends on the objectives of data analysis, i.e., whether we are interested in the point estimation (e.g. image estimation) or in the posterior distribution of parameters of interest (e.g. the tensor probability distribution). If the goal is prone to the former, the augmented data could be fully ignored. For example, in the EM algorithm we are only interested in the mean (conditional expectation) value of the latent variable in order to find the modes of the unknown parameters. From the Bayesian point of view, both the latent variable and the unknown parameters are uncertain and are treated similarly. This means that we need a generating mechanism to calculate the posterior distribution of each unknown. The exposition and advantages of DA provide that the elaboration of specific algorithms are needed in parameter estimation.

7.2 Comparison

We have proposed different methods for tensor estimation in both frequentist and Bayesian framework. Each method has its own merits, the general comparison in statistical viewpoint is given below.

McMC vs VB We conclude this chapter with a further discussion of the differences between the McMC and VB. The aim of McMC is to generate Markov chains on the state space and explore the state space by computing empirical averages under the joint posterior distribution of the unknown parameters and latent variables, while the VB algorithm converges to a fixed point distribution by approximating the posterior marginals. The algorithm is stopped after reaching suitable tolerance(s). These are essential differences between the McMC and the VB methods. Moreover, the success of the MH procedure depends crucially upon the proper choices of the proposals, whilst the VB algorithm intends to speed up computation. But both algorithms need the posterior to be proper. Incidentally, in this chapter we use the terminology: the joint, conditional, and marginal posterior, which had been well-defined in pervasive Bayesian statistics literature, e.g. in Gelman et al. (2014) [Chapter 5].

McMC vs EM In the data augmentation version, the EM algorithm related to the McMC can be seen as a precursor of Gibbs sampler (Robert and Casella , 2004). The EM has, however, essential differences related to the McMC: the McMC is exploring the state space to compute empirical averages under the joint posterior distribution of the unknown parameters and latent variables, whereas the EM in MLE and MAP is a deterministic algorithm. The latter converges to a maximizer of the posterior distribution and is reduced to the MLE when the prior is flat. The augmented data N do not need simulation in these two strategies. While analyzing the joint posterior distribution by McMC is computationally intensive due to all the unknowns including the latent variables, and we need to draw samples from their full conditionals or from the approximated forms. On the other hand, McMC renders full inference about the uncertainty of the unknown parameters via the joint posterior probability distribution. By the analogy, when comparing the frequentist and Bayesian inference, the advantages of the latter is that we can include restrictions to the parameters in forms of probability distributions, e.g. regularization can be simultaneously added into the model in a probabilistic manner. The EM algorithm also works in the Bayesian framework with different perspectives. If our focus is on point estimates, we can invoke the maximum a posteriori (MAP) estimation to achieve the objective by maximizing the posterior density. EM-MAP had been implemented in **PII**.

7.3 Summary of the data and the included papers

7.3.1 Real data

The real data used in this thesis are collected by using either the T2 weighted or the mixed spin echo sequences. The HARDI data used in this thesis are mainly from multiple shells. The authors have no conflicts of interest to disclose.

7.3.2 Summary

The summary of the included papers is given below, where we use the terminology listed at the end of this chapter.

PI Bayesian McMC. Data augmentation in Rician noise model and Bayesian Diffusion Tensor Imaging.

The main contributions:

- Poisson data augmentation, transforming the nonlinear regression model under the Rician noise model into the GLM framework.
- Bayesian modeling and McMC method for diffusion tensor estimation in DTI and HARDI.
- Regularization technique for modeling the tensor dependence by introducing an isotropic prior of the tensor fields by GMRF.

Advantages of this work are:

1. The proposed estimation scheme is under the Rician noise model by data augmentation, which therefore can work on the DW-MRI data in a wide range of the frequency domain.
2. We analyze the probability distribution of each parameter restricted by the prior knowledge.
3. A Bayesian regularization scheme is simultaneously introduced into the model for image denoising.
4. The method is implemented on both synthetic and real data with comparison under the assumptions of tensor independence and dependence.

Shortage:

1. This work does not impose the *positivity constraint* in the tensor estimation.

PII EM in MLE & MAP. Fast Estimation of Diffusion Tensors under Rician noise by the EM algorithm.

The main contributions:

- This work presents a detailed estimation scheme by the EM (greedy) algorithm in MLE and MAP under the 2nd and 4th order tensor models.
- We clarify the difference between the Bayesian and the frequentist method.
- We explain that our EM-MLE is faster than the traditional ML method in theory and in computation.
- We compare our EM algorithm with the EM algorithm in phase space recently presented in the literature, and clarify the difference.
- We apply a stabilized Fisher scoring method for fast convergence of the tensor coefficients.
- We extend our method from the signal compartment model presented in the work to the multicompartment case.
- The method is implemented and the precision is experimented under the 2nd and 4th order tensor models on synthetic and real data.

Advantages of this work are:

1. The proposed scheme is under the Rician noise modeling via the Poisson data augmentation. Therefore, it can work on the DW-MRI data in a wide range of the frequency domain.
2. The method has dramatically reduced the computational burden compared with the traditional MLE method, and can proceed in parallel computation across each voxel.

Shortages:

1. The method is under the assumption of tensor independence without regularization.
2. This work again does not impose the positivity constraints into the tensor parameters.

PIII EM in MLE. An improved EM algorithm for solving MLE in constrained diffusion kurtosis imaging of the human brain.

The main contributions:

- This work introduces von Mises data augmentation, transforming the nonlinear regression model into the GML framework under the Rician noise model in diffusion kurtosis imaging (DKI).
- We propose an EM algorithm in MLE working in DKI, which is one of the advanced diffusion weighting imaging techniques.
- A constrained stabilized Fisher-scoring algorithm is fully presented by using the barrier method, in which the specific constraints in DKI have been imposed into the algorithm, including positivity of the tensor parameters.
- The improvements and accuracy of the estimation scheme have been illustrated by implementing the proposed method on both synthetic and real data by comparing the weighted least squares (WLS) and MLE methods.

Advantages of this work are:

1. The method considers all the necessary constraints in DKI, and it is the fastest among the MLE methods.
2. From model *per se*, DKI can detect the degree of deviation of Gaussianity of diffusion in vivo, providing much more important structural information by the image contrasts.
3. The proposed estimation scheme can work in parallel computation across each voxel, which is therefore practically feasible.

Shortages:

1. The method again is under the assumption of tensor independence without regularization.
2. The method can only work with the b-value less than 3000 mm^2/s, and the data must at least contain three different b-values. These drawbacks are coming from DKI model *per se*.

PIV Bayesian VB. Variational Bayes estimation in constrained kurtosis diffusion imaging under a Rician noise model.

Main contributions:

- We introduce a Bayesian regularization model in DKI, where the tensor dependence is modeled by including an isotropic prior of the quadratic parameters with respect to the tensor coefficients by the GMRF. The positivity constraints in DKI therefore have been imposed directly into the model.
- We implement the Variational Bayes algorithm, which is fully established for DKI estimation by von Mises data augmentation under the Rician noise model.
- A constrained stabilized Fisher scoring method is applied for updating tensor parameters, where we use twice the Laplace approximation and the delta method to construct the estimation scheme via the VB algorithm.
- The method is implemented for both cases (refer to dementia with Lévy bodies) and control data. A test study has been also conducted by using real data.

Advantages of this work are:

1. The proposed scheme has imposed all the natural constraints in DKI.
2. The Bayesian strategy aims for analyzing the posterior probabilities of the parameters of interest. Therefore, the results are expected to be more accurate than the frequent methods. We use the VB algorithm to estimate the optimum of each posterior distribution, which led to computational feasibility in the estimation.

3. Regularization scheme was simultaneously constructed in the modeling under the Bayesian framework, which from the smoothness viewpoint is more efficient and less uncertain compared with the common penalized frequent methods.
4. We analyzed the case and control data of the human brain by the proposed method in this advantage imaging protocol.

Shortages:

1. The proposed scheme can not be applied parallelly across voxels, but it is possible to conduct parallel computation among the blocks of tensors which have certain large distance.
2. Since the work is describing a new method in DKI, it is inevitable that only the DW-MRI data with the b-value less than 3000 mm^2/s and containing at least three different b-values can be considered.

REFERENCES

Aitken, A.C., 1935. Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, 4(02): 106-110.

Alzheimer's Association, 2015. 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: the Journal of the Alzheimer's Association*, 11(3): 332.

Amemiya, T. 1985. *Advanced Econometrics.* Harvard University Press.

Andersen, A.H., 1996. On the Rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 36(2): 331-332.

Andersson, J.L.R., 2008. Maximum a posteriori estimation of diffusion tensor parameters using a Rician noise model: why, how and but. *NeuroImage*, 42(4): 1340-1356.

Assemlal, H.E., Tschumperlé, D. and Brun, L. 2009. Efficient and robust computation of PDF features from diffusion MR signal. *Medical Image Analysis*, 13(5): 715-729.

Atchadé, Y., Fort G., Moulines, E., Priouret, P., 2009. Adaptive Markov chain Monte Carlo: theory and methods. *Preprint*.

Banerjee, S., Carlin, B.P. and Gelfand, A.E., 2004. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Barmpoutis, A., Jian, B., Vemuri, B.C., Shepherd, T.M., 2007. Symmetric positive 4^{th} order tensors & their estimation from diffusion weighted MRI. *Information Processing in Medical Imaging*, 20: 308-319.

Barmpoutis, A., Hwang, M.S., Howland, D., Forder, J.R., 2009. Regularized positive-definite fourth order tensor field estimation from DW-MRI. *NeuroImage*, 45(1): S153-S162.

Barmpoutis, A., Jian, B. and Vemuri, B.C, 2009. Adaptive kernels for multi-fiber reconstruction. *International Conference on Information Processing in Medical Imaging*, Springer, 338-349.

Barmpoutis, A. and Vemuri, B.C., 2009. Groupwise registration and Atlas construction of 4th-order tensor fields using the $\mathbb{R}^+$ Riemannian metric. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, 12(Pt 1): 640-647.

Barmpoutis, A. and Vemuri, B.C., 2010. A unified framework for estimating diffusion tensors of any order with symmetric positive-definite constraints. *2010 IEEE international symposium on biomedical imaging: from nano to macro*, 1385-1388.

Bammer, R., 2003. Basic principles of diffusion-weighted imaging. *European Journal of Radiology*, 45(3): 169-184

Basser, P.J., Mattiello, J., Turner, R., Le Bihan, D., 1993. Diffusion tensor echo-planar imaging of the human brain. *Proceedings of the SMRM*, 584

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

Basser, P.J., Mattiello, J. and Le Bihan, D., 1994b. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1): 259-267.

Basser, P. J., 1995. Inferring microstructural features and the physiological state of tissues from diffusion-weighted images. *NMR in Biomedicine*, 8(7): 333-344.

Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi A., 2000. In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, 44(4): 625-632.

Basser, P.J. and Pajevic, S., 2003. A normal distribution for tensor-valued random variables: applications to diffusion tensor MRI. *IEEE Transactions on Medical Imaging* 22 (7): 785-794.

Basser, P.J. and Pajevic, S., 2007. Spectral decomposition of a 4th-order covariance tensor: Applications to diffusion tensor MRI. *Signal Processing*, 87: 220-236.

Baz, J. and Chacko, G., 2004. *Financial Derivatives: Pricing, Applications, and Mathematics*. Cambridge University Press.

Beal, M.J., Ghahramani, Z. and Rasmussen, C.E., 2002. Advances in neural information processing systems (NIPS 2001). *The Annals of Statistics*, 453-463.

Bernardo, J.M. and Smith, A.F.M., 1994. *Bayesian Theory*. John Wiley & Sons.

Bernstein, M.A., Thomasson, D.M. and Perman, W.H., 1989. Improved detectability in low signal-to-noise ratio magnetic resonance images by means of a phase-corrected real reconstruction. *Medical Physics*, 16(5): 813-817.

Behrens, T.E.J., Woolrich, M.W., Jenkinson, M., Johansen-Berg, H. Nunes, R.G., Clare, S., Matthews, P.M., Brady, J.M., Smith, S.M., 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine*, 50(5): 1077-1088.

Behrens, T.E.J., Berg, H. J., Jbabdi, S., Rushworth, M.F.S., Woolrich, M.W., 2007. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34(1): 144-155.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2): 192-236.

Besag, J., Green, P., Higdon, D., Mengersen, K., 1995. Bayesian computation and stochastic systems. *Statistical Science*, 3-41.

Bloch, F., 1946. Nuclear induction. *Physical Review*, 70(7-8): 460.

Box, G.E.P. and Tiao, G.C., 1992. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons.

Brown, D.F., 1999. Lewy body dementia. *Annals of medicine*, 31(3): 188-196.

Brown, R., and Haacke, M., Cheng, N., Thompson, M., Venkatesan, R., 2004. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. John Wiley & Sons.

Burdette, J.H., Durden, D.D., Elster, A.D., Yen, Y.F., 2001. High b-value diffusion-weighted MRI of normal brain. *JCAT*, 25(4): 515.

Cardenas-Blanco, A., Nezamzadeh, M., Fottit, C., Cameron, I., 2007. Accurate noise bias correction applied to individual pixels. *Proceedings of the International Society of Magnetic Resonance in Medicine*, 3(2): 3445.

Carr, H.Y. and Purcell, E.M., 1954. Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Physical Review*, 94(3): 630-638.

Chandrasekhar, S., 1943. Stochastic problems in physics and astronomy. *Reviews of Modern Physics*, 15(1): 1.

Chung, K.L., 1967. *Markov Chains*. Springer.

Clark, C.A., Hedehus, M. and Moseley, M.E., 2002. In vivo mapping of the fast and slow diffusion tensors in human brain. *Magnetic Resonance in Medicine*, 47(4): 623-628.

Clifford, P., 1990. Markov random fields in statistics. *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, 19-32.

Cramér, H., 1946. *Mathematical Methods of Statistics*. Princeton University Press.

Cory, D.G., 1990. Measurement of translational displacement probabilities by NMR: an indicator of compartmentation. *Magnetic Resonance in Medicine*, 14(3): 435-444.

Cross, G.R. and Jain, A.K., 1983. Markov random field texture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PAMI-5 (1): 25-39.

Das, B.K., 2015. Basic principles of MR imaging. *Positron Emission Tomography*, 185-188.

Damasio, H. 1995. *Human Brain Anatomy in Computerized Images*. Oxford University Press.

Darmois, G., 1935. Sur les lois de probabilitéa estimation exhaustive. *Comptes Rendus de l'Académie des Sciences*, 260: 1265-1266.

Dempster, A. P., Laird, N. M. and Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, S1-S38.

Descoteaux, M., 2010. *High Angular Resolution Diffusion MRI: from Local Estimation to Segmentation and Tractography*. PhD thesis. INRIA Sophia Antipolis, France.

Dong, Q., Welsh, R.C., Chenevert, T.L., Carlos, R.C., Maly-Sundgren, P., Gomez-Hassan, D.M., Mukherji, S.K., 2004. Clinical applications of diffusion tensor imaging. *Journal of Magnetic Resonance Imaging*, 19(1): 6-18.

Durrett, R. 2010. *Probability: Theory and Examples*. Cambridge University Press.

Fan, J.Y., Yuan, Y.X., 2001. On the convergence of a new Levenberg-Marquardt method. *Technical Report, AMSS, Chinese Academy of Sciences*, 1-11.

Firbank, M.J., Blamire, A.M., Krishnan, M.S., Teodorczuk, A., English, P., Gholkar, A., Harrison, R.M., O'Brien, J.T., 2007. Diffusion tensor imaging in dementia with Lewy bodies and Alzheimer's disease. *Psychiatry Research: Neuroimaging*, 155(2): 135-145.

Fisher, R.A., 1925. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5): 700-725.

Fisher, N.I., Lewis, T. and Embleton, B.J.J., 1987. *Statistical Analysis of Spherical Data*. Cambridge University Press.

Frandsen, J., Hobolth, A., Østergaard L, Vestergaard-Poulsen P., Jensen E.B.V., 2007. Bayesian regularization of diffusion tensor images. *Biostatistics*, 8(4): 784-799.

Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The Elements of Statistical Learning*, vol.1. Springer Series in Statistics, Berlin, Springer.

Fukunaga, K., 2013. *Introduction to Statistical Pattern Recognition*. Academic Press.

Gelman, A., Roberts, G., Gilks, W., 1996. Efficient metropolis jumping rules. *Bayesian Statistics*, 5: 599-608.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. *Bayesian Data Analysis*, vol.2. Taylor & Francis.

Geman, S. and Geman, D., 1993. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Journal of Applied Statistics*, 20(5-6): 25-62.

Geman, S. and Graffigne, C., 1986. Markov random field image models and their applications to computer vision. *Proceedings of the International Congress of Mathematicians*, 1(2): 1496-1517.

Ghosh, A., Milne, T. and Deriche, R., 2014. Constrained diffusion kurtosis imaging using ternary quartics & MLE. *Magnetic Resonance in Medicine*, 71(4): 1581-1591.

Giussani, C., Poliakov, A., Ferri, R.T., Plawner, L.L., Browd, S.R., Shaw, D.W.W., Filardi, T.Z., Hoeppner, C., Geyer, J. R., Olson, J.M., Douglas, J.G., Villavicencio, E.H., Ellenbogen, R.G., Ojemann, J.G., 2010. DTI fiber tracking to differentiate demyelinating diseases from diffuse brain stem glioma. *NeuroImage*, 52(1): 217-223.

Goldberger, A.S., 1964. *Econometric Theory*. New York, John Wiley & Sons.

Golumbic, M.C., 2004. *Algorithmic Graph Theory and Perfect Graphs*. Elsevier.

Goss, C.M., 1960. Gray's anatomy of the human body. *Academic Medicine*, 35(1): 90.

Gradshteyn, I.S. and Ryzhik, I.M., 2007. *Table of Integrals, Series, and Products*. Edited by Jeffrey A., Zwillinger D. Academic Press, 918-920.

Green, P.J., 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2): 149-192.

Greene, W.H., 2012. *Econometric Analysis*. Cambridge University Press.

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

Haacke, E.M., Brown, R.W., Thompson, M.R. and Venkatesan R., 1999. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*, vol.82. New York, Wiley Blackwell.

Hahn, E.L., 1950. Spin echoes. *Physical Review*, 80: 580-594.

Hammersley, J.M. and Clifford, P., 1971. Markov fields on finite graphs and lattices. *Unpublished*.

Hans, C., Dobra, A., West, M., 2007. Shotgun stochastic search for "large p" regression. *JASA*, 102(478): 507-516.

Hashemi, R.H., Bradley, W.G. and Lisanti, C.J., 2012 *MRI: the Basics*. Lippincott Williams & Wilkins.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97-109.

Helpern, J.A., Lo, C., Hu, C., Falangola, M.F., Rapalino, O., Jensen, J.H., 2009. Diffusional kurtosis imaging in acute human stroke. *Proceedings 17th Scientific Meeting, International Society for Magnetic Resonance in Medicine*, 3493.

Henkelman, R.M., 1985. Measurement of signal intensities in the presence of noise in MR images. *Medical Physics*, 12(2): 232-233.

Hesselink, J.R. 1996. *Basic Principles of MR Imaging.* Clinical magnetic resonance imaging, 2nd ed. Philadelphia: WB Saunders Company.

Hilbert, D., 1888. Über die darstellung definiter formen als summe von formenquadraten. *Mathematische Annalen*, 32(3): 342-350.

Hobert, J.P. and Casella, G., 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436): 1461-1473.

Hornak, J.P., 1996. *The Basics of MRI*. Rochester Institute of Technology.

Häggström, O., 2002. *Finite Markov Chains and Algorithmic Applications*, 52. Cambridge University Press.

Issidorides, M.R., Mytilineou, C., Panayotacopoulou, M.T., Yahr, M.D., 1991. Lewy bodies in parkinsonism share components with intraneuronal protein bodies of normal brains. *Journal of Neural Transmission-Parkinson's Disease and Dementia Section*, 3(1): 49-61.

Jaakkola, T.S. and Jordan, M.I., 2000. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1): 25-37.

Jensen, J.H. and Helpern, J.A., 2003. Quantifying non-Gaussian water diffusion by means of pulsed-field-gradient MRI. *In: Proceedings of the 11th Annual Meeting of ISMRM*, 2154.

Jensen, J.H., Helpern, J.A., Ramani, A., Lu, H., Kaczynski, K., 2005. Diffusional kurtosis imaging: The quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. *Magnetic Resonance in Medicine*, 53(6): 1432-1440.

Jensen, J.H., Joseph, A., Ramani, A., Lu, H., Kaczynski, K., 2010. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. *NMR in Biomedicine*, 23(7): 698–710.

Jian, B. and Vemuri, B.C., 2007. A unified computational framework for deconvolution to reconstruct multiple fibers from diffusion weighted MRI. *Medical Imaging, IEEE Transactions on*, 26(11): 1464-1471.

Jones, D.K. and Basser, P.J., 2004. "Squashing peanuts and smashing pumpkins": How noise distorts diffusion-weighted MR data. *Magnetic Resonance in Medicine* 52(5): 979-993.

Kaipio, J. and Somersalo, E., 2006. *Statistical and Computational Inverse Problems*, vol.160. Springer.

Kantarci, K., Avula, R., Senjem, M.L., Samikoglu, A.R., Zhang, B., Weigand, S.D., Przybelski, S.A., Edmonson, H.A., Vemuri, P., and Knopman, D.S., 2006. Dementia with Lévy bodies and Alzheimer disease Neurodegenerative patterns characterized by DTI. *Neurology*, 74(22): 1814-1821.

Koay, C.G. and Basser, P.J., 2006. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. *Journal of Magnetic Resonance*, 179(2): 317-322.

Koay, C.G., Chang, L.C., Carew, J.D., Pierpaoli, C., Basser, P.J., 2006. A unifying theoretical and algorithmic framework for least squares methods of estimation in diffusion tensor imaging. *Journal of Magnetic Resonance* , 182(1): 115-125.

Koistinen, P., 2010. *Computational Statistics*. Lecture notes. Department of Mathematics and Statistics, University of Helsinki.

Krissian, K. and Aja-Fernández, S. (2009). Noise-driven anisotropic diffusion filtering of MRI. *IEEE Transactions on Image Processing*, 18(10): 2265-2274.

Kullback, S. and Leibler, R.A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 79-86.

Landman, B., Bazin, P-L. and Prince, J. 2007. Diffusion tensor estimation by maximizing Rician likelihood. *11th IEEE International Conference on Computer Vision ICCV 2007*.

Lauterbur, P.C., 1973. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature*, 242(5394): 190-191.

Lauwers, L., Barbé, K., Van Moer, W., Pintelon, R., 2010. Analyzing Rice distributed functional magnetic resonance imaging data: a Bayesian approach. *Measurement Science & Technology*, 21(11): 115804.

Le Bihan, D., Breton, E., Lallemand, D., Grenier, P., Cabanis, E., Laval-Jeantet, M., 1986. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology*, 161(2): 401-407.

Lu, H., Jensen, J.H, Ramani, A., Helpern, J.A., 2006. Three-dimensional characterization of non-Gaussian water diffusion in humans using diffusion kurtosis imaging. *NMR in Biomedicine*, 19(2): 236-247.

Lu, H., Jensen, J.H., Hu, C., Falangola, M.F., Ramani, A., Ferris, S., Helpern, J.A., 2006. Alterations in cerebral microstructual integrity in normal aging and in Alzheimer's disease: a multi-contrast diffusion MRI study. *Proceedings of the 14th Annual Meeting of ISMRM*, 723.

Mak, E., Su, L., Williams, G.B. and O'Brien, T.J., 2014. Neuroimaging characteristics of dementia with Lévy bodies. *Alzheimer's Research & Therapy*, 6(2): 18.

Mai, J.K., Assheuer, J. and Paxinos, G., 1997. *Atlas of the Human Brain*. Academic Press San Diego.

Maier, S.E., Vajapeyam, S., Mamata, H., Westin, C.F., Jolesz, F.A., Mulkern, R.V., 2004. Biexponential diffusion tensor analysis of human brain diffusion data. *Magnetic Resonance in Medicine*, 51(2): 321-330.

Maier, S.E. and Mulkern, R.V., 2008. Biexponential analysis of diffusion-related signal decay in normal human cortical and deep gray matter. *Magnetic Resonance in Medicine*, 26(7): 897-904.

Mansfield, P., 1977. Multi-planar image formation using NMR spin echoes. *Journal of Physics C: Solid State Physics*, 10(3): L55.

Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(2): 431-441.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 2(6): 1087-1092.

McCullagh, P., Nelder, J.A., 1989. *Generalized Linear Models*, 2nd Edition. Chapman & Hall/CRC.

Mori, S., 2007. *Introduction to Diffusion Tensor Imaging*. Elsevier.

Mori, S. and Van Zijl, P., 1995. Diffusion weighting by the trace of the diffusion tensor within a single scan. *Magnetic Resonance in Medicine*, 33(1): 41-52.

Nelder J.A. and Baker R.J., 1972. *Generalized Linear Models.* Wiley Online Library.

NessAiver, M., Moriel, N.A., 1997. *All you really need to know about MRI physics.* Simply Physics.

Niendorf, T., Dijkhuizen, R.M., Norris, D.G., Van Lookeren Campagne, M., Nicolay, K., 1996. Biexponential diffusion attenuation in various states of brain tissue: Implications for diffusion-weighted imaging. *Magnetic Resonance in Medicine*, 36 (6): 847-857.

NIA and NINDS, 2015. *Lévy Body Dementia: Information for Patients, Families, and Professionals*. NIH Publication.

Nummelin, E., 1984. *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press.

Nummelin, E., 2002. MC's for MCMC'ists. *International Statistical Review*, 70(2): 215-240.

Ormerod, J.T., Wand, M.P., 2010. Explaining variational approximations. *The American Statistician*, 64(2): 140-153.

Papadopoulo, T., Ghosh, A. and Deriche, R., 2007. Complete set of invariants of a 4th order tensor: The 12 tasks of HARDI from ternary quartics. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, 233-240.

Patterson, H.D. and Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3): 545-554.

Pitman, E.J.G., 1936. Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(04): 567-579.

Poot, D. HJ, Arnold, J., Achten, E., Verhoye, M., Sijbers, J., 2010. Optimal experimental design for diffusion kurtosis imaging. *IEEE transactions on medical imaging*, 29(3): 819-829.

Qi L. Yu G. and Wu E.X., 2010. Higher order positive semidefinite diffusion tensor imaging. *SIAM Journal on Imaging Sciences*, 3(3): 416-433.

Rice, S.O., 1944. Mathematical analysis of random noise. *Bell System Technical Journal*, 23(3): 282-332.

Ringman, J.M., O'Neill, J., Geschwind, D., Medina, L., Apostolova, L.G., Rodriguez, Y., Schaffer, B., Varpetian, A., Tseng, B., Ortiz, F., 2007. Diffusion tensor imaging in preclinical and presymptomatic carriers of familial Alzheimer's disease mutations. *Brain*, 130(7): 1767-1776.

Roberts, G., Gelman, A. and Gilks, W., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithm. *Annals of Applied Probability*, 7: 110-120.

Robert, C.P. and Casella, G., 2004. *Monte Carlo Statistical Methods*. Second Edition. Springer Texts in Statistics, New York, Springer.

Roberts, G.O. and Rosenthal, J.S., 2004. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1: 20-71.

Roberts, G. and Rosenthal, J.S., 2001. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4): 351-367.

Rue, H. and Held, L., 2005. *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.

Ruszczyński, A.P. 2006. *Nonlinear Optimization*, vol.2. Princeton University Press.

Salimans, T., and Knowles, D.A. 2013. Fixed-form variational posterior approximation through stochastic linear regression. *International Society for Bayesian Analysis*, 8(4): 837-882.

Salvador, R., Pena, A., Menon, D.K., Carpenter, T.A., Pickard, J.D., Bullmore, E.T., 2004. Formal characterization and extension of the linearized diffusion tensor model. *Human Brain Mapping*, 24(3), 144-155.

Aja-Fernández S. and Vegas-Sanchez-Ferrero G. , 2016. *Statistical analysis of noise in MRI*. Springer

Saul, L.K. and Jordan, M.I., 1996. Exploiting tractable substructures in intractable networks. *Advances in Neural Information Processing Systems*, 486-492.

Schervish, M.J.,1995. *Theory of Statistics*. New York, Springer-Verlag.

Schultz, T. and Seidel, H.P., 2008. Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics*, 14(6): 1635-1642.

Slichter, C., 2013. *Principles of Magnetic Resonance*. Springer Science & Business Media.

Sundberg, R., 1974. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, 49-58.

Sundgren, P.C., Dong, Q., Gomez-Hassan, D., Mukherji, S.K., Maly, P., Welsh, R., 2004. Diffusion tensor imaging of the brain: review of clinical applications. *Neuroradiology*, 46(5): 339-350.

Stejskal, E.O. and Tanner, J.E., 1965. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics*, 42(1): 288-292.

Stieltjes, B., Brunner, R.M., Fritzsche, K., Laun, F., 2013. *Diffusion Tensor Imaging: Introduction and Atlas*. Springer Science & Business Media.

Tabesh, A., Jensen, J.H., Ardekani, B.A., Helpern, J.A., 2011. Estimation of tensors and tensor-derived measures in diffusional kurtosis imaging. *Magnetic Resonance in Medicine*, 65(3): 823-836.

Tanner, M.A. and Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. *JASA*, 82(398): 528-540.

Thomalla, G., Glauche, V. and Weiller, C. Röther J., 2005. Time course of wallerian degeneration after ischaemic stroke revealed by diffusion tensor imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(2): 266-268.

Tierney, L., 1992. Exploring posterior distributions using Markov chains. *DTIC Document*, 563-570.

Tierney, L., 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4): 1701-1728.

Torrey, H., 1956. Bloch equations with diffusion terms. *Physical Review*, 104: 563-565.

Tournier, J., Calamante, F. and Connelly, A., 2012. MRtrix: diffusion tractography in crossing fiber regions. *International Journal of Imaging Systems and Technology*, 22(1), 53-66.

Tuch, D.S., Weisskoff R.M., Belliveau, J.W., Wedeen V.J., 1999. High angular resolution diffusion imaging of the human brain. *Proceedings of the 7th Annual Meeting of ISMRM, Philadelphia*, 321.

Tuch, D.S., 2002. Diffusion MRI of complex tissue structure. Ph.D. thesis, MIT.

Tuch, D.S., Reese, T.G. Wiegell, M.R., Wedeen, V. J., 2003. Diffusion MRI of complex neural architecture. *Neuron*, 40(5): 885-895.

Tutz, G., 2011. *Regression for Categorical Data*, vol.34. Cambridge University Press.

Veraart, J., Van Hecke, W. and Sijbers, J., 2011. Constrained maximum likelihood estimation of the diffusion kurtosis tensor using a Rician noise model. *Magnetic Resonance in Medicine*, 66(3): 678-686.

Watson, R., Blamire, A.M. and Colloby, S.J. Wood, J.S., Barber, R., He, J.B. and O'brien, J.T., 2012. Characterizing dementia with Lewy bodies by means of diffusion tensor imaging. *Neurology*, 79(9): 906-914.

Wozniak, J.R., Mueller, B.A., Bell, C.J., Muetzel, R.L., Lim, K.O., Day, J.W., 2013. Diffusion tensor imaging reveals widespread white matter abnormalities in children and adolescents with myotonic dystrophy type 1. *Journal of Neurology*, 260(4): 1122-1131.

Yablonskiy, D.A., Bretthorst, G. L. and Ackerman Joseph, J.H., 2003. Statistical model for diffusion attenuated MR signal. *Magnetic Resonance in Medicine*, 50(4): 664–669.

Yamashita, N. and Fukushima, M., 2001. On the rate of convergence of the Levenberg-Marquardt method. *Topics in Numerical Analysis*, 239-249.

Zhao, H., 2013. Variational Bayesian Learning and its Applications. Ph.D. thesis. University of Waterloo.

Zhu, H., Zhang, H., Ibrahim, J.G., Peterson, B.S., 2007. Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data. *Journal of the American Statistical Association*, 102(480): 1085-1102.

Zhu, X., Gur, Y., Wang, W., Fletcher, P.T., 2013. Model selection and estimation of multi-compartment models in diffusion MRI with a Rician noise model. *Information Processing in Medical Imaging*, 644-655.

Øksendal, B., 2003. *Stochastic Differential Equations. An Introduction with Applications*. Springer.

Özarslan E. and Mareci T.H., 2003. Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging. *Magnetic Resonance in Medicine*, 50(5): 955-965.

Özarslan, E., Shepherd, T.M., Vemuri, B.C., Blackband, S.J., Mareci, T.H., 2006. Resolution of complex tissue microarchitecture using the diffusion orientation transform (DOT). *NeuroImage*, 31(3): 1086-1103.

Šmídl, V. and Quinn, A., 2006. *The Variational Bayes Method in Signal Processing*. Springer Science & Business Media.

ORIGINAL PAPERS

I

DATA AUGMENTATION IN RICIAN NOISE MODEL AND BAYESIAN DIFFUSION TENSOR IMAGING

by

Gasbarra, D, Liu, J & Railavo, J. 2019

Submitted manuscript

Data augmentation in Rician noise model and Bayesian Diffusion Tensor Imaging

DARIO GASBARRA *, JIA LIU † AND JUHA RAILAVO ‡

Abstract

Diffusion Magnetic Resonance Imaging is a powerful technique for detecting anisotropies in the diffusion of water molecules that corresponds to nervous fibers in the living brain. In this process, spectral data from the displacement distribution of water molecules are collected by a magnetic resonance scanner. The signals are corrupted by a Rician noise, which leads to a non-linear regression problem. Diffusion tensor imaging is the simplest approach postulating a Gaussian displacement distribution at each volume element. The common inference is based on the linearized log-normal regression model that can only fit the spectral data at low frequencies. This solution, however, fails to treat with the high frequency data containing detailed information of the water displacement with low signal to noise ratios. In this paper, we propose to use Poisson data augmentation to represent the Rician likelihood, and directly work with the Rician noise model and cover the full spectral range. We propose a Bayesian hierarchical model and a Markov chain Monte Carlo method in performance of tensor estimation with the 2nd and 4th of tensor models in diffusion MRI. A regularization scheme is suggested in Bayesian framework for image smoothness. The method has implemented with real data of human brain.

Key words and phrases: Brain Imaging, Bayesian Smoothing, Data Augmentation, Generalized Linear Model, Gaussian Random Field, Image Regularization, Inverse Problem, Markov Chain Monte Carlo, Poissonization.

1 Introduction

Diffusion as a physical phenomenon has been an essential part of the history and development of magnetic resonance imaging. Hahn (1950) observed the effect of diffusion to spin-echoes, Carr and Purcell (1954) studied the effects of diffusion on free precession,

*Department of Mathematics and Statistics, University of Helsinki P.O. Box 68 FI-00014 Finland e-mail: dario.gasbarra@helsinki.fi

†Corresponding author, Department of Mathematics and Statistics, University of Helsinki P.O. Box 68 FI-00014 Finland; Department of Mathematics and Statistics, University of Jyväskylä, P.O. Box 35 (MaD) FI-40014 Finland e-mail:jia.liu@helsinki.fi; jia.2.liu@jyu.fi

‡HUS e-mail:juha.railavo@elisanet.fi

and Torrey (1956) modified the Bloch equations to include diffusion term with spatially varying magnetic field. Stejskal and Tanner (1965), in their seminal paper, introduced the pulsed gradient spin echo sequence and showed the potential of diffusion related signal attenuation to probe the motion of molecules and to define the diffusion coefficient. In 1973 P. Lauterbur (who shared the Nobel Prize with Sir Peter Mansfield in 2003) made history publishing his groundbreaking paper entitled "Image formation by induced local interactions: Examples employing nuclear magnetic resonance ". In his experiment Lauterbur superimposed a magnetic field gradient on the static uniform magnetic field. Because of the Larmor principle, different parts of the sample would have different resonance frequencies and so a given resonance frequency could be associated with a given position. He also pointed out that it is possible to measure molecular diffusion from the decay of the MR-signal. Diffusion weighted magnetic resonance imaging was introduced by Le Bihan et al. (1986), measuring the displacement of protons. Moseley et al. (1990) observed that diffusion in the white matter was anisotropic. In anisotropic media the mobility of the molecules is orientation dependent and can not be represented by one single diffusion coefficient. The three dimensional process of diffusion modeled by diffusion tensors was introduced by Basser et al. (1994). In neuroimaging, we measure restricted diffusion within neuron cells, and the principal diffusion eigenvector corresponds to the direction of a nervous fiber.

The Rician distribution is the law of the square root of a non-central χ^2 random variable. It appears in several applied fields, including signal processing, and also in mathematical finance as the transition density of the Bessel process, modeling stochastic volatility and short rate (Baldeaux and Platen, 2013). It is well known that the noise in an MR measurement has a Rician distribution instead of Gaussian (Jones and Basser, 2004; Henkelman, 1985; Zhu et al., 2007; Assemlal et al., 2009; Landman et al., 2007). Several authors e.g., Zhu et al. (2007); Salvador et al. (2005) add the noise-induced bias into the measurement so that a simple Gaussian noise model can be fitted to the data. But none of them can easily gain the potential important information (e.g., Mori and Tournier, 2014; Burdette et al., 2001) from the high-frequency data, because in the high b-value range the corrected data does not fit the Gaussian distribution. The Rician noise model was used in e.g., Gudbjartsson and Patz (1995); Andersson (2008); Lauwers et al. (2010), but in all these cases the methods dealing with Rician noise are computationally intensive. In this work we will deal directly with the Rician likelihood by using data augmentation, reducing the non-standard regression problem to the standard Poisson regression. This novel approach applies to the full spectral range, including the observations in the low SNR regime which after discretization are recorded as zeros. In Liu et al. (2016) the EM algorithm based on the same data augmentation was used to compute the Maximum Likelihood (MLE) and Maximum A Posteriori (MAP) estimators. Here we follow the Bayesian approach and, after building a hierarchical model assigning an isotropic Gaussian prior to the diffusion tensor, we use Markov chain Monte Carlo to analyze the posterior distribution. The proposed method applies directly to diffusivity models of increasing complexity as higher order tensor models. Another

major difference compared with Liu et al. (2016) is that in this work we also model the voxel dependence for the image regularization.

The paper is structured as follows: the nonlinear regression problem with Rician noise model is described in Section 2. The main contribution of the paper, data augmentation by Poissonization is introduced in Section 2.2. In Section 3, after a general discussion on McMC methods, we construct the Bayesian hierarchical model for a single tensor (Section 3.2), and the Gibbs-Metropolis algorithm for sampling posterior distribution (Section 3.3). In Sections 4.4 and 4.5, we reformulate different tensor models into our Bayesian framework. Section 5.1 conduct simulation studies to compare the performance of our method and the other popular methods in DTI from several synthetic datasets. The implementation of these methods is illustrated in Section 5.2 with an analysis of human brain data.

2 Generalized linear modeling with Rician likelihood

2.1 Rician noise

We shall consider signal-observation pairs (S, Y) with $S \in \mathbb{R}$ and

$$Y = |S + \varepsilon_x + i\varepsilon_y| = \sqrt{(S + \varepsilon_x)^2 + \varepsilon_y^2}\,, \tag{2.1}$$

where $(\varepsilon_x, \varepsilon_y)$ are independent with Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and $\varepsilon = (\varepsilon_x + i\varepsilon_y)$ is a complex Gaussian noise. It follows that Y has a Rician distribution with density

$$p_{S,\sigma^2}(y) = \frac{y}{\sigma^2} \exp\left(-\frac{y^2 + S^2}{2\sigma^2}\right) I_0\left(\frac{yS}{\sigma^2}\right), \tag{2.2}$$

where

$$I_0(z) = \frac{1}{\pi} \int_0^\pi \exp(z \cos t) dt \tag{2.3}$$

is the modified Bessel function of first kind. When the signal to noise ratio (SNR) $|S|/\sigma$ is large enough, the Rician likelihood Equation (2.2) is well approximated by a log-normal density with mean $\log(S)$ and variance $\sigma^2 S^{-2}$. In such case S and σ^2 are estimated by using iterated Weighted Least Squares (WLS) (see Zhu et al., 2007; Koay et al., 2006). We will consider instead the low SNR regime and work directly with the Rician likelihood.

2.2 Poissonization and data augmentation

Lemma 2.1. *Consider random variables (N, X), where N is Poisson distributed with mean $t > 0$, and given N, X has a conditional distribution Gamma$(N + 1, 1/(2\sigma^2))$, that is*

$$\begin{aligned} &P_{t,\sigma^2}(N = n, X \in dx) = P_t(N = n) P_{\sigma^2}(X \in dx | N = n) \\ &= \frac{(tx)^n}{(n!)^2 (2\sigma^2)^{n+1}} \exp\left(-t - \frac{x}{2\sigma^2}\right) dx\,. \end{aligned} \tag{2.4}$$

Then

1. *$Y := \sqrt{X}$ has a Rician marginal with density*

$$P_{t,\sigma^2}(Y \in dy) = \frac{y}{\sigma^2} \exp\left(-t - \frac{y^2}{2\sigma^2}\right) I_0\left(\frac{y}{\sigma}\sqrt{2t}\right) dy.$$

2. *The conditional distribution of N given Y is given by*

$$P_{t,\sigma^2}(N = n|Y = y) = I_0\left(\frac{y}{\sigma}\sqrt{2t}\right)^{-1} \left(\frac{y^2 t}{2\sigma^2}\right)^n (n!)^{-2}. \tag{2.5}$$

In particular $P_{t,\sigma^2}(N = 0|Y = 0) = 1$.

Proof 1. *1 is well known. After a change of variable sum over n by using the representation*

$$I_0(2z) = {}_0F_1(1, z^2) = \sum_{n=0}^{\infty} \frac{z^{2n}}{(n!)^2} \tag{2.6}$$

(Gradshteyn and Ryzhik, 2015), where ${}_0F_1(1, z)$ is a Gaussian hypergeometric function. Equation (2.5) is a consequence of the Bayes formula.

We shall give a name to the distribution Equation (2.5). In Appendix Equation (A) we discuss random sampling from it.

Definition 2.2. *For $\tau > 0$, consider two i.i.d. random variables N, N' with Poisson(τ) distribution, and define the probability distribution*

$$p_\tau(n) := P_\tau(N = n|N = N') = I_0(2\tau)^{-1} \frac{\tau^{2n}}{(n!)^2}, \quad n \in \mathbb{N}.$$

We call $(p_\tau(n) : n \in \mathbb{N})$ the reinforced Poisson distribution with parameter τ.

2.3 Non-linear regression and reduction to GLM

In the follow-up we assume a non-linear regression model with Rician noise and signals $S_i = \exp(Z_i\theta)$, $i = 1, \ldots, m$, where $Z \in m \times (d + 1)$ is a known design matrix, while $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^\top \in \mathbb{R}^{d+1}$ and σ^2 are the unknown parameters.

From a statistician's point of view, a non-linear regression problem is most conveniently framed in the context of *Generalized Linear Models* (GLM), where the measurements have probability density of the form

$$p_{Z\theta,\phi}(y) = f_{\tau,\phi}(y) = c(y, \phi) \exp\left(\frac{y\tau - a(\tau)}{\phi}\right), \tag{2.7}$$

see McCullagh and Nelder (1989). The function $a(\tau)$ in Equation (2.7) specifies an exponential family of distributions for the response Y, and τ is determined implicitly by the relation $g(\mu) = Z\theta$, where $\mu = E_{\tau,\phi}(Y) = a'(\tau)$ is the expectation and $g(\mu)$ is the *link* function. Unfortunately, this assumption is not satisfied by the Rician likelihood in Equation (2.2). In order to reduce the non-linear regression problem to the framework of generalized linear models, by using Equation (2.4), we propose a novel data

augmentation strategy for parameter estimation under the exact Rician likelihood. For each data point (Y_i, Z_i), we introduce an unobservable variable N_i which follows a GLM with Poisson response corresponding to $a(\tau) = \exp(\tau)$, $\phi = 1$, and link function $g(\mu) = \log(2\sigma^2\mu)/2$, and by Lemma 2.1 we obtain

Corollary 2.3. *In the settings of Lemma 2.1 with $t = \exp(2Z\theta)/(2\sigma^2)$ we have the following.*

- *The marginal distribution of Y has a Rician density of Equation (2.2).*
- *The conditional distribution $P_t(N = n|Y = y)$ is a reinforced Poisson distribution $p_\tau(n)$ with parameter*

$$\tau = \frac{y \exp(Z\theta)}{2\sigma^2} \ .$$

3 Bayesian Inference

3.1 Use of improper priors

We discuss first the integrability of the Rician likelihood 2.2 with respect to a flat improper prior for $\theta = (\theta_0, \theta_1, \ldots, \theta_d) \in \mathbb{R}^{d+1}$.

Proposition 3.1. *The following conditions are equivalent.*

1. $$\int_{\mathbb{R}^{d+1}} \prod_{j=1}^{m} \left\{ \frac{Y_j}{\sigma^2} \exp\left(-\frac{Y_j^2 + \exp(2Z_j\theta)}{2\sigma^2} \right) I_0\left(\frac{Y_j \exp(Z_j\theta)}{\sigma^2} \right) \right\} d\theta < \infty.$$

2. $$\inf_{\theta \in \mathbb{R}^{(d+1)}} \max_{1 \le j \le m} \left\{ Z_j\theta \right\} > 0 \ . \tag{3.8}$$

3. *The convex cone generated by the Z-rows covers the whole space $\mathbb{R}^{d+1}$.*

Proof. Directly from the inequalities

$$\exp\left(-\frac{\big(Y - S\big)^2}{2\sigma^2} \right) \ge \exp\left(-\frac{Y^2 + S^2}{2\sigma^2} \right) I_0\left(\frac{YS}{\sigma^2} \right) \ge \exp\left(-\frac{\big(Y + S\big)^2}{2\sigma^2} \right),$$

$$\text{for } Y, S \ge 0 \ . \quad \square \tag{3.9}$$

The problem arises from the signal model $S_j = \exp(Z_j\theta)$. With a linear model $S_j = Z_j\theta$, the Rician likelihood would be always integrable under the flat prior $d\theta$.

Consider now the improper prior $\pi(d\sigma) = \sigma^{-2} d\sigma^2$ for the noise variance. We have

$$\int_0^\infty \prod_{j=1}^{m} \left\{ \frac{Y_j}{\sigma^2} \exp\left(-\frac{Y_j + \exp(Z_j\theta)}{2\sigma^2} \right) I_0\left(\frac{Y_j \exp(Z_j\theta)}{\sigma^2} \right) \right\} \sigma^{-2} d\sigma^2 \le$$

$$\prod_{j=1}^{m} Y_j \int_0^\infty \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{m} \big(Y_j - \exp(Z_j\theta)\big)^2 \right) \sigma^{-2(1+m)} d\sigma^2 =$$

$$\left(\sum_{j=1}^{m} \big(Y_j - \exp(Z_j\theta)\big)^2 \right)^{-m} (m-1)! 2^m \prod_{j=1}^{m} Y_j,$$

which is bounded w.r.t θ when the linear system $Z\theta = \log Y$ does not have solution, as it is the case with probability 1 when $\text{rank}(Z) < m$. In such case the posterior distribution will be proper when the prior of θ is proper, and also with a flat θ-prior under condition Equation (3.8).

3.2 Prior Specification and Hierarchical Model

In the follow-up we shall consider the $(d+1)$-dimensional non-linear regression model with design matrix $[\mathbf{1}^\top Z]$ and parameter $(\theta_0, \theta_1, \ldots, \theta_d)^\top$ producing the signals $S_i = S_0 \exp(Z_i\theta)$, $i = 1, \ldots, m$, with $S_0 > 0$. The intercept $\theta_0 = \log S_0$ plays a special role and it is assigned a conjugate prior. Namely,

- S_0^2 has a Gamma prior with shape and rate parameters $a_S, b_S > 0$, respectively.
- $\theta \in \mathbb{R}^d$ has a Gaussian prior with precision matrix Ω and mean $\mu_\theta \in \mathbb{R}^d$.
- σ^2 has an inverse Gamma prior with shape and rate parameters $a_\sigma, b_\sigma \geq 0$, respectively. We include also the case with $a_\sigma = b_\sigma = 0$, which corresponds to the scale invariant prior with density $\pi(\sigma^2) \propto 1/\sigma^2$. We have shown in 3.1 that such an improper prior combined with the likelihood produces a proper posterior distribution.

Given the parameters (θ, S_0, σ^2), the augmented data pairs $\{(N_i, X_i) : i = 1, \ldots, m\}$ are conditionally independent with conditional distributions

- $\big[N_i \big| \theta, S_0, \sigma^2\big] \sim \text{Poisson}\Big(S_0^2 \exp(2\theta \cdot Z_i)/(2\sigma^2)\Big)$,
- $\big[X_i \big| N_i, \sigma^2\big] \sim \text{Gamma}\big(N_i + 1, 1/\big(2\sigma^2\big)\big)$, and $\quad Y_i = \sqrt{X_i}$.

3.3 Gibbs-Metropolis algorithm

We describe in details the Markov chain Monte Carlo (McMC) algorithm sampling efficiently the parameters and the augmented data from the posterior distribution $p(\theta, S_0, \sigma^2, N|Y)$. For the general theory of McMC we refer to Robert and Casella (2005). We combine sequentially several block updates, where in turn a subset of parameters is updated keeping the remaining ones fixed. When it is feasible, we sample the parameters from their full conditional distribution (Gibbs' update). For the regression parameter θ, we use Laplace approximation to construct a Gaussian proposal distribution approximating the full conditional.

Updating N The auxiliary random variables N_i are updated by sampling from the full conditional distribution. Conditionally on θ, S_0, σ^2 and the measurements (Y_i, Z_i), the r.v.'s N_i are conditionally independent reinforced Poisson distributed with parameters

$$\tau_i = \frac{Y_i \exp(Z_i\theta) S_0}{2\sigma^2}\ ,\ i = 1, \ldots, m,$$

respectively. In Appendix A we discuss Monte Carlo sampling from the reinforced Poisson distribution.

Remark 3.2. *The augmented data N is generated "on the fly" from the full conditional distribution above when needed. It is not necessary to store N into the computer memory.*

Updating θ Conditionally on $N = (N_i : i = 1, \ldots, m)$, S_0^2 and σ^2, the parameter θ is independent of the observations Y_i, the full conditional distribution being proportional to

$$p(\theta|\sigma^2, N) \propto \exp\left(-\frac{1}{2}(\theta - \mu_\theta)^\top \Omega(\theta - \mu_\theta) + \left(2\sum_{i=1}^m N_i Z_i\right)\theta - \frac{S_0^2}{2\sigma^2}\sum_{i=1}^m \exp(2Z_i\theta)\right). \tag{3.10}$$

We choose a Gibbs-Metropolis update with Gaussian proposal distribution

$$q(\theta|\hat{\theta}) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^\top \big(\Omega + J(\hat{\theta})\big)(\theta - \hat{\theta})\right), \tag{3.11}$$

where have employed the Laplace approximation of Equation (3.10) around its mode $\hat{\theta}$. Here σ^2 and N are fixed and the precision matrix is the Fisher information

$$J(\theta) = E_\theta\left(\nabla_\theta \log p(N|\theta, S_0\sigma^2)^\top \nabla_\theta \log p(N|\theta, S_0\sigma^2)\right) = \frac{2S_0^2}{\sigma^2}\sum_{i=1}^m \exp(2Z_i\theta) Z_i^\top Z_i.$$

To find the mode $\hat{\theta}$, we use the iterative Fisher scoring algorithm (see McCullagh and Nelder, 1989; Lange, 2013, Chapter 10), and the details can be found in Liu et al. (2016). The Hastings log-ratio for $\widetilde{\theta}$ sampled from the proposal distribution $q(\cdot|\hat{\theta})$ is given by

$$\log\left(\frac{p(\widetilde{\theta}|\sigma^2, N)q(\theta|\hat{\theta})}{p(\theta|\sigma^2, N)q(\widetilde{\theta}|\hat{\theta})}\right) = \left(\hat{\theta}^\top\big(\Omega + J(\hat{\theta})\big) - \mu_\theta^\top\Omega - 2\sum_{i=1}^m N_i Z_i\right)(\theta - \widetilde{\theta})$$
$$+\frac{S_0^2}{2\sigma^2}\sum_{i=1}^m\big\{\exp(2Z_i\theta) - \exp(2Z_i\widetilde{\theta})\big\} + \frac{1}{2}\widetilde{\theta}^\top J(\hat{\theta})\widetilde{\theta} - \frac{1}{2}\theta^\top J(\hat{\theta})\theta\ .$$

Remark 3.3. *Computing the Laplace approximation Equation (3.11) to the full conditional density Equation (3.10), is crucial in order to get high acceptance rates in McMC. Without data augmentation, the GLM-likelihood in Equation (3.10) should be replaced by a product of Rician likelihoods. It is also possible to compute by Fisher scoring the Laplace approximation of the full conditional under such Rician likelihood. However, for large sample size m, it could be not computationally affordable to do that at every McMC update of every single tensor.*

The algorithm is based on the assumption that the Fisher scoring algorithm converges to same global maximum $\hat{\theta}$ for all initial values θ. However, with a finite number of iterations, the approximate mode $\breve{\theta}$ obtained by starting the Fisher scoring algorithm from the proposal value $\widetilde{\theta}$ will be slightly different than the approximate mode $\hat{\theta}$ obtained starting from the initial value θ. In order to correct for this discrepancy we have to run the Fisher scoring algorithm a second time starting from the proposed value $\widetilde{\theta}$

and reaching another approximate maximum $\check{\theta}$. In this case we redefine the Hastings log-ratio as

$$\log\left(\frac{p(\widetilde{\theta}|\sigma^2,N)q(\theta|\check{\theta})}{p(\theta|\sigma^2,N)q(\widetilde{\theta}|\hat{\theta})}\right) = \frac{1}{2}\log\left(\frac{\det(\Omega+J(\check{\theta}))}{\det(\Omega+J(\hat{\theta}))}\right)$$
$$+\left(\mu_\theta^\top\Omega + 2\sum_{i=1}^m N_i Z_i\right)(\widetilde{\theta}-\theta) + \frac{S_0^2}{2\sigma^2}\sum_{i=1}^m\{\exp(2Z_i\theta)-\exp(2Z_i\widetilde{\theta})\}$$
$$-\frac{1}{2}\widetilde{\theta}^\top\Omega\widetilde{\theta}+\frac{1}{2}\theta\Omega\theta+\frac{1}{2}(\widetilde{\theta}-\hat{\theta})^\top\big(\Omega_0+J(\hat{\theta})\big)(\widetilde{\theta}-\hat{\theta})-\frac{1}{2}(\theta-\check{\theta})^\top\big(\Omega+J(\check{\theta})\big)(\theta-\check{\theta})\ .$$

Updating S_0 We see that

$$p(N|\theta,S_0,\sigma^2)\propto(S_0^2)^a\exp\big(-bS_0^2\big)$$

with

$$a=\sum_{i=1}^m N_i,\quad b=\frac{1}{2\sigma^2}\sum_{i=1}^m\exp\big(2Z_i\theta\big)\ ,$$

and the conjugate prior $\pi(S_0^2)\propto\text{Gamma}(a_S,b_S)$. Then S_0^2 is Gamma(a_S+a,b_S+b)-distributed conditionally on θ,N and σ^2. We sample ξ from this Gamma distribution and set $S_0=\sqrt{\xi}$. For small values of the shape parameter we sample $\log(S_0)$ by using the rejection sampling algorithm of Liu et al. (2017), described in Appendix B.

Updating S_0 **without data augmentation** We derive an alternative proposal distribution for S_0 by substituting in the full conditional distribution $[S_0|\theta,\sigma^2,N,Y]$ the augmented data (N_i) with the current values of the conditional expectations $E(N_i|Y_i,S_0,\theta,\sigma^2)$. For fixed σ^2 and θ, define

$$\tau_i(S_0)=\frac{S_0\exp(Z_i\theta)Y_i}{2\sigma^2}\ \text{ and }\quad w(\tau)=\tau\frac{I_1(2\tau)}{I_0(2\tau)}.$$

The proposal is obtained by taking the square root of $\widetilde{S}_0^2$ sampled from the gamma distribution

$$q(S_0^2\to\widetilde{S}_0^2)\propto(\widetilde{S}_0^2)^{\left(a_S-1+\sum_{i=1}^m w(\tau_i(S_0))\right)}\exp\left(-\widetilde{S}_0^2\left(b_S+\frac{1}{2\sigma^2}\sum_{i=1}^m\exp(2Z_i\theta)\right)\right).$$

The Hastings log-ratio for the transition $S_0\to\widetilde{S}_0$ is given by

$$\sum_{i=1}^m\left\{\log I_0\big(2\tau_i(\widetilde{S}_0)\big)-\log I_0\big(2\tau_i(S_0)\big)\right\}$$
$$+\log\Gamma\left(a_S+\sum_{i=1}^m W(\tau_i(S_0))\right)-\log\Gamma\left(a_S+\sum_{i=1}^m W(\tau_i(\widetilde{S}_0))\right)$$
$$+\left\{2\log(S_0)+\log\left(b_S+\frac{1}{2\sigma^2}\sum_{i=1}^m\exp(2Z_i\theta)\right)\right\}\sum_{i=1}^m w(\tau_i(\widetilde{S}_0))$$
$$-\left\{2\log(\widetilde{S}_0)+\log\left(b_S+\frac{1}{2\sigma^2}\sum_{i=1}^m\exp(2Z_i\theta)\right)\right\}\sum_{i=1}^m w(\tau_i(S_0))\ .$$

Updating σ^2 The variance parameter is updated in the Gibbs step. Conditionally on the augmented data (N_i, Y_i, Z_i) and the parameter θ, the conditional density of σ^2 up to a multiplicative constant is given by

$$p(\sigma^2|\theta, S_0, N_i, Y_i, Z_i, i = 1, \dots, m) \propto$$
$$\exp\left(-\frac{1}{\sigma^2}\left\{b_\sigma + \frac{1}{2}\sum_{i=1}^{m}\left\{Y_i^2 + \exp(2Z_i\theta)S_0^2\right\}\right\}\right)(\sigma^2)^{-\left(a_\sigma+1+\sum_{i=1}^{m}(2N_i+1)\right)},$$

which corresponds to the inverse gamma distribution with shape and rate parameters

$$\left(a_\sigma + \sum_{i=1}^{m}(2N_i+1)\right) \quad \text{and} \quad \left(b_\sigma + \frac{1}{2}\sum_{i=1}^{m}\left(Y_i^2 + \exp(2Z_i\theta)S_0^2\right)\right), \text{ respectively.}$$

Remark 3.4. *Note that the noise variance σ^2 appears in both augmented likelihood factors*

$$p(N_i|Z, \theta, S_0, \sigma^2)p(Y_i|N_i, \sigma^2),$$

which makes the pair (S_0, σ^2) identifiable.

Updating σ^2 without data augmentation Alternatively, for fixed $S_0 = S_0(v)$ and $\theta = \theta(v)$, let $S_i = S_0\exp(Z_i\theta)$, and let

$$W_i(\sigma^2) = \frac{I_1\left(Y_iS_i\sigma^{-2}\right)}{I_0\left(Y_iS_i\sigma^{-2}\right)}.$$

Then we propose $\widetilde{\sigma}^2$ from the inverse Gamma proposal distribution

$$q(\sigma^2 \to \widetilde{\sigma}^2) \propto \widetilde{\sigma}^{-2(m+a_\sigma+1)}\exp\left(-\frac{1}{\widetilde{\sigma}^2}\left(b_\sigma + \sum_{i=1}^{m}\left\{\frac{Y_i^2+S_i^2}{2} - Y_iS_iW_i(\sigma^2)\right\}\right)\right),$$

with the Hastings log-ratio for accepting the move $\sigma^2 \to \widetilde{\sigma}^2$ given by

$$(m+a_\sigma)\left\{\log\left(b_\sigma + \sum_{i=1}^{m}\left\{\frac{Y_i^2+S_i^2}{2} - Y_iS_iW_i(\widetilde{\sigma}^2)\right\}\right)\right.$$
$$\left. - \log\left(b_\sigma + \sum_{i=1}^{m}\left\{\frac{Y_i^2+S_i^2}{2} - Y_iS_iW_i(\sigma^2)\right\}\right)\right\} + \sigma_0^{-2}\sum_{i=1}^{m}Y_iS_iW_i(\widetilde{\sigma}^2)$$
$$- \widetilde{\sigma}^{-2}\sum_{i=1}^{m}Y_iS_iW_i(\sigma^2) + \sum_{i=1}^{m}\left\{\log I_0(Y_iS_i\widetilde{\sigma}^{-2}) - \log I_0\left(Y_iS_i\sigma^{-2}\right)\right\}.$$

4 Diffusion Tensor Imaging

In the follow-up we apply the method to Diffusion Tensor Imaging (DTI), with a special emphasis on the choice of the rotation invariant Gaussian priors for the tensor parameter θ.

4.1 The 2nd-order Tensor Model

Without going into the physics of DTI, we sketch the diffusion data acquisition from the statistical point of view. After applying two consecutive and opposite gradient pulses with amplitude $|\mathbf{q}|$ in the direction $\mathbf{u} = \mathbf{q}/|\mathbf{q}| \in \mathcal{S}^2$,* with time delay t, MR produces at every spatial location v a signal

$$S_v(\mathbf{q}) \;=\; S_v(\mathbf{0}) E_v\Big(\exp\big(i\ \mathbf{q}\cdot\mathbf{V}_t\big)\Big) \;= \\ S_v(\mathbf{0})\exp\Big(-\frac{1}{2}\ \mathbf{q}D_v\mathbf{q}^\top\Big) = S_v(\mathbf{0})\exp\Big(-b\mathbf{u}D_v\mathbf{u}^\top\Big), \tag{4.12}$$

where $S_v(\mathbf{0})$ is the concentration of water molecules at v, the control $\mathbf{q}$ is a 3-dimensional pulse gradient, and $b = |\mathbf{q}|^2/2$. In Equation (4.12) appears the characteristic function of a centered Gaussian random vector $\mathbf{V}_t$ with covariance matrix D_v†, which is interpreted as the displacement of a water molecule with initial position v in the time interval $[0, t]$ between the two pulses. The symmetric and positive definite matrix-valued field (D_v) describes the geometry of the media and it is the object of interest. Note that for an eigenvector $\mathbf{q}$ with eigenvalue $g > 0$ satisfying $D\mathbf{q} = g\mathbf{q}$, the MR signal

$$S(\mathbf{q}) = S(\mathbf{0})\exp\big(-\frac{g}{2}|\mathbf{q}|^2\big) \tag{4.13}$$

is highest when $\mathbf{q}$ belongs to the eigenspace of the smallest eigenvalue of D, and lowest in the principal direction. In neuroimaging, we measure restricted diffusion within neuron cells, and the principal diffusion eigenvector corresponds to the direction of the neurons in the voxel v. The MR-signals $S_v(\mathbf{q})$ are measured with additive complex Gaussian noise Equation (2.1), and magnitude measurements $Y_v(\mathbf{q})$ are recorded. It is assumed that these are independent and Rician distributed with signal and noise parameters $S_v(\mathbf{q}_i)$ and σ_v^2, respectively. Since MR-images may contain local artefacts, it is safer to assume that the noise level varies with the spatial location v. Note also that the MR-data is digitalized. Small values of the measurements, which are possible at high b-value, are coded as zeros. In order to use the log-normal approximation and to estimate the parameters by WLS, these zero values should be discarded, inducing sampling bias. With our method we don't need to do that, simply the latent variables take value $N_i = 0$ when $Y_i = 0$ (Lemma 2.1) and contribute to the augmented likelihood accordingly.

In this simple 2nd-order tensor model it is convenient to parametrize the signal as $S(\mathbf{q}) = S_0\exp(Z\theta)$, where

$$\theta = (\theta_1, \ldots, \theta_d)^\top := \big(D_{xx}, D_{yy}, D_{zz}, D_{xy}, D_{xz}, D_{yz}\big)^\top$$

is the vector of tensor parameters, and the design matrix Z has rows

$$Z(\mathbf{q}) = -\big(\mathbf{q}_x^2/2, \mathbf{q}_y^2/2, \mathbf{q}_z^2/2, \mathbf{q}_x\mathbf{q}_y, \mathbf{q}_x\mathbf{q}_z, \mathbf{q}_y\mathbf{q}_z\big)\ .$$

*$\mathcal{S}^2 \subset \mathbb{R}^3$ denotes the unit sphere.

†In the neuroimaging literature another convention is used, where $D = E\big(\mathbf{V}_t^\top\mathbf{V}_t\big)/2$ is referred as *diffusion tensor* and $b = |\mathbf{q}|^2$.

Isotropic Gaussian Prior for the 2nd-Order Tensors Note that Equation (3.8) is never satisfied in the DTI experiment: to see that, consider $\theta = (\theta_0, \theta_D) \in \mathbb{R}^{d+1}$ such that $\theta_D \in \mathbb{R}^d$ parametrizes a positive 2nd-order tensor D and

$$\log(S_0) = \theta_0 < \frac{1}{2} \min\{\mathbf{q}_j D \mathbf{q}_j^\top : 1 \leq j \leq m\} .$$

Therefore, in order to obtain a proper posterior distribution, it is necessary to use a proper prior distribution for both the tensor parameter θ_D and intensity S_0. In DTI it is natural to assign a diffusion tensor prior which is invariant under rotations of the coordinate system. In Basser and Pajevic (2003), Jeffreys (1962), Gasbarra et al. (2017), it is shown that the distribution of zero mean 3×3 symmetric Gaussian random matrix $D = (D_{i,j} : 1 \leq i \leq j \leq 3)$ is isotropic if and only if it has a density of the form

$$p(D) = \frac{\eta^{5/2}\sqrt{\eta + 3\lambda}}{(\pi\sqrt{2})^3} \exp\left(-\frac{1}{2}\left(\eta \text{Trace}(D^2) + \lambda\{\text{Trace}(D)\}^2\right)\right) \tag{4.14}$$

with $\eta > 0$ and $\lambda > -\eta/3$, and Equation (4.14) follows from the characterization of an isotropic Gaussian random field in terms of the distribution of its spherical harmonic coefficients.

For the vector $\theta = (D_{11}, D_{22}, D_{33}, D_{12}, D_{13}, D_{23})$, this corresponds to a Gaussian distribution with zero mean and precision matrix

$$\Omega_D = \begin{pmatrix} \lambda+\eta & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+\eta & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+\eta & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\eta & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\eta & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\eta \end{pmatrix} . \tag{4.15}$$

In particular $E(D_{ij}^2) = (2\eta)^{-1}$ when $i \neq j$, and $E\big(D_{ii}D_{jj}\big) = \big(\delta_{ij} - \lambda/(\eta + 3\lambda)\big)\eta^{-1}$, with negative correlations for $\lambda > 0$.

4.2 Modeling diffusivity with the 4th-order tensors

Several authors, Basser and Pajevic (2007); Mori and Tournier (2014); Ghosh et al. (2009); Moakher (2009); Ghosh et al. (2012) argue that the 2nd-order tensor model Equation (4.12) fails to capture complex tissue structures such as fibers crossing and branching in a single voxel. In fact, while we have a diffusion matrix at every spatial location in the time scales we are considering, the scale of water diffusion is of smaller order than the spatial resolution of the image. The 2nd-order tensor model assumes that the diffusion matrix is constant within one voxel. In reality a voxel contains a whole population of cellular structures, corresponding to a population of diffusion tensors. Instead of measuring the characteristic function of a centered Gaussian random vector, the MR-experiment measures the characteristic function of a Gaussian mixture, and

consequently Equation (4.12) should be replaced by

$$\frac{S_v(\mathbf{q})}{S_v(\mathbf{0})} \;=\; E_v\Big(\exp\big(i\;\mathbf{q}\cdot\mathbf{V}_t\big)\Big) \;=\; \int\limits_{\mathcal{M}^+} \exp\Big(-\frac{1}{2}\mathbf{q}^\top D\mathbf{q}\Big)dQ_v(D)\;, \tag{4.16}$$

which is the characteristic function of the random displacement $\mathbf{V}_t$ of a water molecule randomly selected within the voxel. Here Q_v is a probability distribution for the population of diffusion tensors living in the space $\mathcal{M}^+ \subset \mathbb{R}^{6\times 6}$ of symmetric and positive definite matrices. We see from Equation (4.16) that the signal $S_v(\mathbf{q})$ must be a decreasing w.r.t. $|\mathbf{q}|$. In order to model the exponential decay, we introduce the *diffusivity*

$$d_v(u) = \frac{2\big(\log S_v(0) - \log S_v(\mathbf{q})\big)}{|q|^2},$$

where $u = \mathbf{q}/|\mathbf{q}|$ is the gradient direction. In the 4-th order tensor model it is assumed that the signals are given by

$$S(\mathbf{q}) = S_0\exp\big(-bd(\mathbf{u})\big) = S_0\exp\big(Z\theta\big), \qquad \mathbf{q}\in\mathbb{R}^3, \tag{4.17}$$

with diffusivity

$$d(\mathbf{u}) = D:(\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u}\otimes\mathbf{u}) := \sum_{i_1=1}^{3}\sum_{i_2=1}^{3}\sum_{i_3=1}^{3}\sum_{i_4=1}^{3} D_{i_1i_2i_3i_4}u_{i_1}u_{i_2}u_{i_3}u_{i_4}, \quad \mathbf{u}\in S^2, \tag{4.18}$$

a homogeneous polynomial of degree 4, parametrized by the totally symmetric 4-th order tensor

$$D = \big(D_{i_1i_2i_3i_4} : 1\le i_1\le i_2\le i_3\le i_4\le 4\big)\;.$$

In the left-hand side of Equation (4.17) the tensor parameter are given by

$$\begin{aligned}\theta = \big(&D_{1111}, D_{2222}, D_{3333}, D_{1122}, D_{1133}, D_{2233},\\ &D_{1123}, D_{1223}, D_{1233}, D_{1112}, D_{1113}, D_{1222}, D_{2223}, D_{1333}, D_{2333}\big)^\top,\end{aligned}$$

and the design matrix $Z\in\mathbb{R}^{m\times 15}$ has rows

$$\begin{aligned}-\,(&u_1^4, u_2^4, u_3^4, 6u_1^2u_2^2, 6u_1^2u_3^2, 6u_2^2u_3^2,\\ &12u_1^2u_2u_3, 12u_2^2u_1u_3, 12u_3^2u_1u_2, 4u_1^3u_2, 4u_1^3u_3, 4u_2^3u_1, 4u_2^3u_3, 4u_3^3u_1, 4u_3^3u_2)b.\end{aligned}$$

Isotropic Gaussian Prior for the 4th-order Tensors In Basser and Pajevic (2007), the 4th-order tensor in dimension 3 is shown to be isomorphic to a 2nd-order tensor in dimension 6 under the isomorphism

$$D \longmapsto \widehat{D} := \begin{pmatrix} D_{1111} & D_{1122} & D_{1133} & \sqrt{2}D_{1112} & \sqrt{2}D_{1113} & \sqrt{2}D_{1123}\\ D_{1122} & D_{2222} & D_{2233} & \sqrt{2}D_{1222} & \sqrt{2}D_{1223} & \sqrt{2}D_{2223}\\ D_{1133} & D_{2233} & D_{3333} & \sqrt{2}D_{1233} & \sqrt{2}D_{1333} & \sqrt{2}D_{2333}\\ \sqrt{2}D_{1112} & \sqrt{2}D_{1222} & \sqrt{2}D_{1233} & 2D_{1122} & 2D_{1123} & 2D_{1223}\\ \sqrt{2}D_{1113} & \sqrt{2}D_{1223} & \sqrt{2}D_{1333} & 2D_{1123} & 2D_{1133} & 2D_{1233}\\ \sqrt{2}D_{1123} & \sqrt{2}D_{2223} & \sqrt{2}D_{2333} & 2D_{1223} & 2D_{1233} & 2D_{2233}\end{pmatrix}. \tag{4.19}$$

The six eigenvalues and eigentensors of the 4-th order tensor D, correspond to the eigenvalues and eigenvectors of the matrix $\widehat{D}$. Furthermore, it is shown in Ghosh et al. (2012), that $\text{Trace}(\widehat{D})^2, \text{Trace}(\widehat{D}^2)$ and the polynomial

$$\begin{aligned} g(D) &= D_{1111}(D_{2222}+D_{3333}) + D_{2222}D_{3333} + 3\Big\{D_{1122}^2 + D_{1133}^2 + D_{2233}^2\Big\} \\ &+ 2\Big\{D_{1122}D_{3333} + D_{1133}D_{2222} + D_{2233}D_{1111} + D_{1122}(D_{1133}+D_{2233}) + D_{2233}D_{1133}\Big\} \\ &+ 4\Big\{D_{1233}(D_{1233}-D_{1222}-D_{1112}) + D_{1223}(D_{1223}-D_{1113}-D_{1333}) \\ &+ D_{1123}(D_{1123}-D_{2333}-D_{2223}) - D_{1222}D_{1112} - D_{1113}D_{1333} - D_{2223}D_{2333}\Big\} \end{aligned} \tag{4.20}$$

are invariant under 3D-rotations and span the space of isotropic homogeneous polynomials of degree 2 in the variables D. Here we give the general form of a zero-mean isotropic Gaussian distribution for the 4th-order tensor, with density

$$\begin{aligned} \pi(D) = 2^3\sqrt{\frac{(\gamma+\eta)^9(3\eta-4\gamma)^5(3\eta+8\gamma+15\lambda)}{\pi^{15}}}\exp\Bigg(-\frac{1}{2}\Big\{\eta\text{Trace}(\widehat{D}^2) \\ + \lambda\,\text{Trace}(\widehat{D})^2 + \gamma g(D)\Big\}\Bigg)\,. \end{aligned} \tag{4.21}$$

Again Equation (4.21) follows from the characterization of isotropic Gaussian random fields in terms of the law of their spherical harmonic coefficients, see e.g., Barmpoutis et al. (2009); Özarslan and Mareci (2003).

Under Equation (4.21), the random coefficients $(D_{1111}, D_{2222}, D_{3333}, D_{1122}, D_{1133}, D_{2233})$ are Gaussian zero mean and precision matrix

$$\Omega' = \begin{pmatrix} \eta+\lambda & \lambda+\gamma & \lambda+\gamma & 2\lambda & 2\lambda & 2\lambda+2\gamma \\ \lambda+\gamma & \eta+\lambda & \lambda+\gamma & 2\lambda & 2\lambda+2\gamma & 2\lambda \\ \lambda+\gamma & \lambda+\gamma & \eta+\lambda & 2\lambda+2\gamma & 2\lambda & 2\lambda \\ 2\lambda & 2\lambda & 2\lambda+2\gamma & 6\eta+6\gamma+4\lambda & 4\lambda+2\gamma & 4\lambda+2\gamma \\ 2\lambda & 2\lambda+2\gamma & 2\lambda & 4\lambda+2\gamma & 6\eta+6\gamma+4\lambda & 4\lambda+2\gamma \\ 2\lambda+2\gamma & 2\lambda & 2\lambda & 4\lambda+2\gamma & 4\lambda+2\gamma & 6\eta+6\gamma+4\lambda \end{pmatrix}, \tag{4.22}$$

and are independent from $(D_{1112}, D_{1113}, D_{1222}, D_{2223}, D_{1333}, D_{2333}D_{1123}, D_{1223}, D_{1233})$,

which are Gaussian with zero mean and precision matrix

$$\Omega'' = \begin{pmatrix} 4\eta & 0 & -4\gamma & 0 & 0 & 0 & 0 & 0 & -4\gamma \\ 0 & 4\eta & 0 & 0 & -4\gamma & 0 & 0 & -4\gamma & 0 \\ -4\gamma & 0 & 4\eta & 0 & 0 & 0 & 0 & 0 & -4\gamma \\ 0 & 0 & 0 & 4\eta & 0 & -4\gamma & -4\gamma & 0 & 0 \\ 0 & -4\gamma & 0 & 0 & 4\eta & 0 & 0 & -4\gamma & 0 \\ 0 & 0 & 0 & -4\gamma & 0 & 4\eta & -4\gamma & 0 & 0 \\ 0 & 0 & 0 & -4\gamma & 0 & -4\gamma & 12\eta + 8\gamma & 0 & 0 \\ 0 & -4\gamma & 0 & 0 & -4\gamma & 0 & 0 & 12\eta + 8\gamma & 0 \\ -4\gamma & 0 & -4\gamma & 0 & 0 & 0 & 0 & 0 & 12\eta + 8\gamma \end{pmatrix}. \tag{4.23}$$

The covariance matrix of D is positive definite under the constraints

$$\eta > 0, \quad \frac{3}{4}\eta > \gamma > -\eta, \quad \lambda > -\left(\frac{1}{5}\eta + \frac{8}{15}\gamma\right).$$

Positivity constraint for the 4th-order tensors. Because the diffusivity function models signal decay, the 4-th order tensor must satisfy the positivity constraint

$$D : (\mathbf{u} \otimes \mathbf{u} \otimes \mathbf{u} \otimes \mathbf{u}) \geq 0, \quad \forall \mathbf{u} \in \mathcal{S}^2 ,$$

implying that the 6×6 matrix $\widehat{D}$ in Equation (4.19) has positive eigenvalues. This is a sufficient but not a necessary condition: it is enough to have positivity on the algebraic variety

$$\left\{(u_1^2, u_2^2, u_3^2, u_1u_2, u_1u_3, u_2u_3) : \ (u_1, u_2, u_3) \in \mathbb{R}^3\right\} \subset \mathbb{R}^6 .$$

When $\widehat{D}$ is negative definite, we should check the sign of the Z-eigenvalue of the diffusivity, which was introduced by Qi et al. (2010) as the solution of the constrained optimization problem

$$\lambda = \min\left\{d(\mathbf{u}) : \ \mathbf{u} \in \mathbb{R}^3, \ |\mathbf{u}| = 1\right\} .$$

4.3 Positivity constraints in McMC

In general, there are two simple ways to include a constraint $C \subset \mathbb{R}^d$ in a McMC algorithm. In order to approximate the constrained expectation

$$E_\pi(g(\xi)|\xi \in C) = \frac{E_\pi(g(\xi)\mathbf{1}(\xi \in C))}{\pi(C)} = \frac{\int_{\mathbb{R}^d} g(x)\mathbf{1}_C(x)f(x)dx}{\int_{\mathbb{R}^d} \mathbf{1}_C(x)f(x)dx},$$

one has to choose:

- include the constraint into the target distribution obtaining a new target density proportional to $\widetilde{f}(x) = f(x)\mathbf{1}_C(x)$. In practice this means starting from a state $\xi_0 \in C$, and rejecting every proposed state which does not satisfy the constraint. The resulting Markov chain takes values in the constraint set C.

- alternatively, include the constraint in the test function and sample from the unconstrained Metropolis-Hastings algorithm. By the law of large numbers, with probability 1

$$E_\pi(g(\xi)|\xi \in C) = \lim_{T\to\infty} \frac{\sum_{t=0}^{T} g(\xi_t)\mathbf{1}_C(\xi_t)}{\sum_{t=0}^{T} \mathbf{1}_C(\xi_t)}.$$

The second method has the advantage of simplicity, it is not even required to start the Markov chain from $\xi_0 \in C$, and the unconstrained Markov chain may have better mixing properties than the constrained one. The drawback is that the samples not satisfying the constraint are lost.

4.4 Bayesian regularization of the tensor field

Bayesian regularization is an image-denoising technique, introduced by Geman (1984), which has been already applied in DTI studies (Frandsen et al., 2007). It is assumed that under the prior distribution the spatial parameters of the model are not independent but form a correlated random field. This is a reasonable assumption in our context: even when a priori we do not have any information about the main tensor direction at a given voxel, we know that often tensors from neighbour voxels are similar, just because a nervous fiber possibly continues from one voxel to the next. The prior dependence is taken into account according to Bayes formula and it has a smoothing and denoising effect on the posterior estimates. An alternative is to estimate first the parameters independently at each voxel, and then interpolate the preliminary tensor estimators to obtain a smoothed estimator. The advantage of Bayesian regularization is that estimation and regularization are performed in a single procedure by using all the available information.

We construct a proper isotropic Gaussian prior for a Markov random field of (3×3) symmetric matrices $(D(v) : v \in V)$ where V is the set of voxels, provided with the

neighbourhood relation $v \sim w$ in the $\mathbb{Z}^3$ lattice. Define the (proper) prior density

$$
\begin{aligned}
&\pi(D(v) : v \in V) = \\
&(2\pi)^{-|V|d/2} \det(\Omega)^{|V|/2} \det(I_V + \rho L_V)^{d/2} \exp\Big(-\frac{1}{2} \theta^\top \{ \ (I_V + \rho L_V) \otimes \Omega \} \theta \Big) \\
&\propto \exp\Big(-\frac{1}{2} \sum_{v \in V} \Big\{ \eta \mathsf{Trace}(D(v)^2) + \lambda \{ \ \mathsf{Trace}(D(v)) \big)^2 \} - \\
&\frac{\rho}{2} \sum_{v \sim w} \Big(\eta \mathsf{Trace}(\{D(v) - D(w)\}^2) + \lambda \big\{ \mathsf{Trace}(D(v) - D(w)) \big\}^2 \Big) \Big) \\
&= \exp\Big(-\rho \sum_{v \sim w} \sum_{i=1}^{3} \Big\{ \frac{(\eta + \lambda)}{2} (D_{ii}(v) - D_{ii}(w))^2 \\
&+ \sum_{j<i} \Big(\lambda (D_{ii}(v) - D_{ii}(w))(D_{jj}(v) - D_{jj}(w)) + \eta (D_{ij}(v) - D_{ij}(w))^2 \Big) \Big\} \\
&- \sum_{v \in V} \sum_{i=1}^{3} \Big\{ \frac{(\eta + \lambda)}{2} D_{ii}(v)^2 + \sum_{j<i} \Big(\lambda D_{ii}(v) D_{jj}(v) + \eta (D_{ij}(v))^2 \Big) \Big\} \Big)
\end{aligned}
\tag{4.24}
$$

with hyperparameters $\eta \geq 0, \ \lambda > -\eta/3$ and $\rho \geq 0$, which tune the dependence between tensors at different voxels. L_V denotes the Laplacian matrix of the graph V.

As in Section 2.2, for each voxel v we introduce:

- A regression parameter vector

$$
\begin{aligned}
\theta(v) &= \big(\theta_1(v), \theta_2(v), \theta_3(v), \theta_4(v), \theta_5(v), \theta_6(v) \big) \big) \\
&= (D_{11}(v), D_{22}(v), D_{33}(v), D_{12}(v), D_{13}(v), D_{23}(v) \big).
\end{aligned}
$$

- An independent intensity parameter $S_0(v)$ with $S_v(0)^2 \sim \text{Gamma}(c_1, c_2)$.
- A noise parameter $\sigma^2(v) > 0$ with scale-invariant improper prior $\propto \big(\sigma^2(v) \big)^{-1}$ and
- a random vector $N(v) = (N_k(v) : k = 1, \ldots, m)$ which follows the generalized linear model of Corollary 2.3 with Poisson response distribution and logarithmic link function, covariate matrix $Z \in m \times d$ and parameters $\theta(v), S_0(v), \sigma^2(v)$.

Here $(\sigma(v) : v \in V)$ are independent and $(N(v) : v \in V)$ are conditionally independent given $(\theta(v) : v \in V)$.

As before, we compute the Laplace approximation for the log-likelihood at each voxel v. When we combine this Gaussian log-likelihood approximation with the pairwise-difference Gaussian prior by using Bayes formula, we obtain an approximating Gaussian posterior for $\theta(v)$, which we will use as the proposal distribution in the Gibbs-Metropolis update. We may consider the single site update, where $\theta(v)$ is updated voxelwise conditionally on $N(v)$ and the values $\theta(w)$ at neighbour voxels $v \sim w$. Alternatively we can construct a Gaussian approximation to the full conditional as a joint proposal in a simultaneous update for a block $(\theta(v) \in W)$, where $W \subseteq V$ is a connected

subset of voxels. The size of a block can vary from a single site to the whole brain. For example, we may define a block as a ball with given center and radius under the graph distance, which is the length of the shortest path between two voxels. We denote the exterior boundary of W by

$$\partial W := \{w \in V \setminus W : \ \exists v \in W \text{ with } w \sim v\}$$

and set $\overline{W} := W \cup \partial W$, $\partial\{v\} := \{w \in V : w \sim v\}$ denotes the neighbourhood of v, and $\deg(v) = \#\partial\{v\}$ is the degree of v. We update the variable $(\theta(w) : w \in W)$ conditional on the observations $(N(w) : w \in W)$ and $(\theta(v) : v \in \partial W)$.

The prior of $(\theta(w) : w \in W \cup \partial W)$ is Gaussian and the likelihood of $\theta(w)$ with respect to the augmented data $N(w)$ is approximated by the Gaussian density $\mathcal{N}(\hat{\theta}(w), \hat{J}(w)^{-1})$, where $\hat{\theta}(w)$ and $\hat{J}(w)$ are functions of $N(w), S_0(w), \sigma^2(w)$ and the design matrix Z, computed by using Fisher scoring under the Poisson GLM as in Section 3.3. The corresponding Gaussian posterior distribution $q(\theta(w) : w \in W)$ will be used as a proposal in the Metropolis block update, and satisfies

$$\begin{aligned}
&\log q(\theta(w) : w \in W) = \\
&\text{const} - \frac{\rho}{2} \sum_{w\sim v : v\in W, w\in\overline{W}} \Big(\eta \text{Trace}(\{D(v) - D(w)\}^2) + \lambda \big\{\text{Trace}(D(v) - D(w))\big\}^2 \Big) \\
&- \frac{1}{2} \sum_{v\in W} \Big\{ (\theta(v) - \hat{\theta}(v))^\top \hat{J}(v)(\theta(v) - \hat{\theta}(v)) + \theta(v)^\top \Omega \theta(v) \Big\} \\
= \text{const} &- \frac{1}{2} \sum_{v\in W} \theta(v)^\top \Big(\big(1 + \deg(v)\rho\big)\Omega + \hat{J}(v) \Big) \theta(v) + \rho \sum_{v\sim w : v,w\in W} \theta(v)^\top \Omega \theta(w) \\
&+ \sum_{v\in W} \theta(v)^\top \Big(\hat{J}(v)\hat{\theta}(v) + \rho\Omega \Big(\sum_{w\in\partial\{v\}\setminus W} \theta(w) \Big) \Big) \\
= \text{const} &- \frac{1}{2} \sum_{v,w\in W:\ w=v \text{ or } w\sim v} (\theta(v) - \hat{\mu}(v))^\top \ \hat{\Psi}_{v,w} \ (\theta(w) - \hat{\mu}(w)) \ ,
\end{aligned}$$

where the constant term does not depend on $(\theta(v) : v \in W)$ and may change from line to line, and after completing the squares we have defined

$$\mu^\top = (\hat{\Psi})^{-1} \hat{\xi}^\top \quad \text{with} \quad \hat{\xi}(v)^\top = \hat{J}(v)\hat{\theta}(v) + \rho\Omega \Big(\sum_{w\in\partial\{v\}\setminus W} \theta(w) \Big) \quad \text{and}$$

$$\hat{\Psi}_{v,w} = \Big(\deg(v)\mathbf{1}(v = w) - \mathbf{1}(v \sim w) \Big) \rho\Omega + \mathbf{1}(v = w)\big(\Omega + \hat{J}(v)\big)$$

that is a band diagonal precision matrix with $(d \times d)$ blocks and $v, w \in W$. This corresponds to a Gaussian proposal distribution $q(\theta(w) : w \in W)$ with mean $(\hat{\mu}(w) : w \in W)$ and covariance $(\hat{\Psi})^{-1}$.

Prior contribution The prior contribution is derived as the proposal contribution by conditioning on the values $(\theta(v) : v \in \partial W)$ without including data. We obtain

$$\begin{aligned}
&\log \pi(\theta(w) : w \in W; \theta(v), v \in \partial W) = \text{const} \\
&-\frac{\rho}{2} \sum_{v\sim w : v\in W, w\in \overline{W}} (\theta(v)-\theta(w))^\top \Omega(\theta(v)-\theta(w)) - \frac{1}{2}\sum_{v\in W} \theta(v)^\top \Omega \theta(v) \\
&= \text{const} - \frac{1}{2}\sum_{v,w\in W} \theta(v)^\top \Phi_{v,w}\theta(w) + \rho \sum_{v\in W} \theta(v)^\top \Omega\Big(\sum_{w\in \partial\{v\}\setminus W} \theta(w)\Big)
\end{aligned}$$

$$\text{with} \qquad \Phi_{v,w} := \Big(\big(1+\deg(v)\rho\big)\mathbf{1}(v=w) - \rho \mathbf{1}(v\sim w)\Big)\Omega, \quad v,w \in W.$$

These expressions determine the Hastings ratio for this Gibbs-Metropolis update (here omitted).

4.5 Updating the regularization parameters of the 2nd-order tensor field

The precision matrix of the Gaussian random field $(\theta(v) : v \in V)$ is the Kronecker product $(I_V + \rho L_V) \otimes \Omega$, where

$$L_V(v,w) = \deg(v)\mathbf{1}(v=w) - \mathbf{1}(v\sim w)$$

denotes the *Laplacian matrix* of the graph $(V, \sim)$ (see Lovász and Vesztergombi (2002)), and Ω was given in Equation (4.15). Since

$$\det((I_V + \rho L_V)\otimes\Omega) = \det(I_V + \rho L_V)^d \det(\Omega)^{|V|},$$

the likelihood for λ, η based on $\big(\theta(v) : v \in V\big)$ is proportional to

$$\begin{aligned}
&\big(\eta^{5/2}\sqrt{\eta+3\lambda}\big)^{|V|} \exp\Big(-\frac{\rho}{2}\sum_{v\sim w}\Big(\eta \text{Trace}(\{D(v)-D(w)\}^2) \\
&+ \lambda\big\{\text{Trace}(D(v)-D(w))\big\}^2 - \frac{1}{2}\sum_{v\in V}\Big(\eta\text{Trace}\big(D(v)^2\big) + \lambda\big\{\text{Trace}(D(v))\big\}^2\Big),
\end{aligned}$$

with constraints $\eta > 0$ and $\lambda > -\eta/3$, $\rho \geq 0$.

In order to factorize the likelihood we reparametrize with $\delta = (\eta + 3\lambda)$, obtaining

$$\begin{aligned}
&\eta^{|V|5/2}\exp\Big(-\eta\Big\{\rho\sum_{v\sim w}\Big(\frac{1}{2}\text{Trace}(\{D(v)-D(w)\}^2) - \frac{1}{6}\big\{\text{Trace}(D(v)-D(w))\big\}^2\Big) + \\
&+ \sum_{v\in V}\Big(\frac{1}{2}\text{Trace}(D(v)^2) - \frac{1}{6}\big\{\text{Trace}(D(v))\big\}^2\Big)\Big\}\Big) \\
&\times\ \delta^{|V|/2}\exp\Big(-\frac{\delta}{6}\Big(\rho\sum_{v\sim w}\big\{\text{Trace}(D(v)-D(w))\big\}^2 + \sum_{v\in V}\big\{\text{Trace}(D(v))\big\}^2\Big)\Big).
\end{aligned}$$

Assuming independent gamma priors for η, δ,

$$\pi(\eta) \sim \text{Gamma}(c_1', c_2'), \quad \pi(\delta) \sim \text{Gamma}(c_1'', c_2''),$$

we obtain the full conditional distribution of (δ, η) as the product of two Gamma densities,

$$\pi(\delta|\theta) \sim$$
$$\text{Gamma}\Big(c_1'' + \frac{|V|}{2}, c_2'' + \frac{\rho}{6}\sum_{v\sim w}\big\{\text{Trace}(D(v) - D(w))\big\}^2 + \frac{1}{6}\sum_{v\in V}\big\{\text{Trace}(D(v))\big\}^2\Big),$$
$$\pi(\eta|\theta) \sim \text{Gamma}\Big(c_1' + \frac{|V|5}{2},$$
$$c_2' + \rho\sum_{v\sim w}\Big(\frac{1}{2}\text{Trace}(\{D(v) - D(w)\}^2) - \frac{1}{6}\big\{\text{Trace}(D(v) - D(w))\big\}^2\Big) +$$
$$\sum_{v\in V}\Big(\frac{1}{2}\text{Trace}(D(v)^2) - \frac{1}{6}\big\{\text{Trace}(D(v))\big\}^2\Big)\Big) .$$

In McMC, we update the regularization parameters by sampling (η, δ) independently from these full conditional distribution and setting $\lambda = (\delta - \eta)/3$.

4.6 Updating the parameters of the 4th-order tensor field

The likelihood for λ, η, γ based on $\big(\theta(v) : v \in V\big)$ is proportional to

$$\mathbf{1}\big(\eta > 0\big)\mathbf{1}\big(3/4\eta > \gamma > -\eta\big)\mathbf{1}\big(\lambda + \eta/5 + \gamma 8/15 > 0\big)$$
$$\Big\{(\gamma+\eta)^9(3\eta - 4\gamma)^5(3\eta + 8\gamma + 15\lambda)\Big\}^{|V|/2}\exp\Big(-\frac{\rho}{2}\sum_{v\sim w}\Big(\eta\text{Trace}(\{\widehat{D}(v) - \widehat{D}(w)\}^2)$$
$$+ \lambda\big\{\text{Trace}(\widehat{D}(v) - \widehat{D}(w))\big\}^2 + \gamma g\big(D(v) - D(w)\big)\Big)$$
$$- \frac{1}{2}\sum_{v\in V}\Big(\eta\text{Trace}(\widehat{D}(v)^2) + \lambda\big\{\text{Trace}(\widehat{D}(v))\big\}^2 + \gamma g(D(v))\Big)\Big) ,$$

where the polynomial $g(D)$ was given in Equation (4.20). In order to factorize the likelihood we reparametrize it as

$$\alpha = (\gamma + \eta),\ \beta = (3\eta - 4\gamma),\ \delta = (3\eta + 8\gamma + 15\lambda)$$

with $\alpha, \beta, \delta > 0$. The linear system has a solution

$$\eta = \frac{\beta + 4\alpha}{7},\ \lambda = \frac{7\delta + 5\beta - 36\alpha}{105},\ \gamma = \frac{3\alpha - \beta}{7}, \tag{4.25}$$

and the corresponding likelihood is proportional to

$$\alpha^{|V|9/2}\exp\Bigg(-\frac{\alpha\rho}{14}\sum_{v\sim w}\Big\{4\text{Trace}(\{\widehat{D}(v)-\widehat{D}(w)\}^2-\frac{12}{5}\big\{\text{Trace}(\widehat{D}(v)-\widehat{D}(w))\big\}^2$$
$$+3g\big(D(v)-D(w)\big)\Big\}-\frac{\alpha}{14}\sum_{v\in V}\Big\{4\text{Trace}(\widehat{D}(v)^2)-\frac{12}{5}\big\{\text{Trace}(\widehat{D}(v))\big\}^2+3g\big(D(v)\big)\Big\}\Bigg)$$
$$\times\,\beta^{|V|5/2}\exp\Bigg(-\frac{\beta\rho}{14}\sum_{v\sim w}\Big\{\text{Trace}(\{\widehat{D}(v)-\widehat{D}(w)\}^2)+\frac{1}{3}\big\{\text{Trace}(\widehat{D}(v)-\widehat{D}(w))\big\}^2$$
$$-g\big(D(v)-D(w)\big)\Big\}-\frac{\beta}{14}\sum_{v\in V}\Big\{\text{Trace}(\widehat{D}(v)^2)+\frac{1}{3}\big\{\text{Trace}(\widehat{D}(v))\big\}^2-g\big(D(v)\big)\Big\}\Bigg)\times$$
$$\delta^{|V|/2}\exp\Bigg(-\frac{\delta\rho}{30}\sum_{v\sim w}\big\{\text{Trace}(\widehat{D}(v)-\widehat{D}(w))\big\}^2-\frac{\delta}{30}\sum_{v\in V}\big\{\text{Trace}(\widehat{D}(v))\big\}^2\Bigg).$$

We assume independent gamma priors for α,β,δ,

$$\pi(\alpha)\sim\text{Gamma}(c_1,c_2),\ ,\pi(\beta)\sim\text{Gamma}(c_1',c_2'),\pi(\delta)\sim\text{Gamma}(c_1'',c_2''),$$

and obtain the full conditional distribution of (α,β,δ) as the product of these Gamma densities:

$$\pi(\alpha|\theta)\sim\text{Gamma}\Bigg(c_1+\frac{9}{2}|V|,c_2+\frac{\rho}{14}\sum_{v\sim w}\Big\{4\text{Trace}(\{\widehat{D}(v)-\widehat{D}(w)\}^2-$$
$$\frac{12}{5}\big\{\text{Trace}(\widehat{D}(v)-\widehat{D}(w))\big\}^2+3g\big(D(v)-D(w)\big)\Big\}\Bigg)$$
$$+\frac{1}{14}\sum_{v\in V}\Big\{4\text{Trace}(\widehat{D}(v)^2-\frac{12}{5}\big\{\text{Trace}(\widehat{D}(v))\big\}^2+3g\big(D(v)\big)\Big\}\Bigg)\Bigg)$$
$$\pi(\beta|\theta)\sim\text{Gamma}\Bigg(c_1'+\frac{5}{2}|V|,c_2'+\frac{\rho}{14}\sum_{v\sim w}\Big\{\text{Trace}(\{\widehat{D}(v)-\widehat{D}(w)\}^2)+$$
$$\frac{1}{3}\big\{\text{Trace}(\widehat{D}(v)-\widehat{D}(w))\big\}^2-g\big(D(v)-D(w)\big)\Big\}\Bigg)$$
$$+\frac{1}{14}\sum_{v\in V}\Big\{\text{Trace}(\widehat{D}(v)^2)+\frac{1}{3}\big\{\text{Trace}(\widehat{D}(v))\big\}^2-g\big(D(v)\big)\Big\}\Bigg)\Bigg)$$
$$\pi(\delta|\theta)\sim\text{Gamma}\Bigg(c_1{}''\frac{|V|}{2},c_2{}''+$$
$$\frac{\rho}{30}\sum_{v\sim w}\big\{\text{Trace}(\widehat{D}(v)-\widehat{D}(w))\big\}^2+\frac{1}{30}\sum_{v\in V}\big\{\text{Trace}(\widehat{D}(v))\big\}^2\Bigg).$$

In the McMC algorithm, (α,β,δ) are updated independently by sampling from these full conditionals. The corresponding parameters (η,λ,γ) are then obtained from Equation (4.25).

When the diffusivity function is assigned voxelwise as

$$d_v(u)=\sum_{\ell=0}^{n}\sum_{m=-2\ell}^{2\ell}\theta_{2\ell,m}(v)Y_{2\ell,m}(u),\quad v\in V,u\in S^2,$$

with common truncation level n, we define the (improper) regularization prior for the random field by assigning a Gaussian prior to the coefficients' pairwise differences as follows:

$$\pi\big(\theta_{2\ell,m}(v) : 0 \le \ell \le n, -2\ell \le m \le 2\ell\big) \propto \prod_{\ell=0}^{n} a_{2\ell}^{-(4\ell+1)|V|}$$
$$\exp\bigg(-\frac{1}{2}\sum_{\ell=0}^{n} a_{2\ell}^{-2} \sum_{m=-2\ell}^{2\ell} \bigg\{ \rho \sum_{v\sim w} \big\{\theta_{2\ell,m}(v) - \theta_{2\ell,m}(w)\big\}^2 + \sum_{v\in V} \theta_{2\ell,m}(v)^2 \bigg\}\bigg).$$

The Bayesian computations of Sections 3.3 and 4.4 apply directly with parameter

$$\theta(v) = \big(\ \theta_{2\ell,m}(v) : 0 \le \ell \le n, -2\ell \le m \le 2\ell\big)^\top \in \mathbb{R}^d,\ d = (2n+1)(n+1),$$

design matrix $Z \in \mathbb{R}^{m\times d}$ with rows

$$Z(\mathbf{q}) = \big(-bY_{2\ell,m}(u) : 0 \le \ell \le n, -2\ell \le m \le 2\ell\big), \quad u = \mathbf{q}/|\mathbf{q}|,\ b = |\mathbf{q}|^2/2\ ,$$

and diagonal precision matrix $\Omega \in \mathbb{R}^{d\times d}$ with diagonal entries

$$\big(a_0^{-2}, a_2^{-2}, a_2^{-2}, a_2^{-2}, a_2^{-2}, a_2^{-2}, \ldots, \underbrace{a_{2n}^{-2}, \ldots, a_{2n}^{-2}}_{(4n+1)\text{ times}}\big)\ .$$

Assuming an independent Gamma prior for the angular power spectrum coefficients, given as

$$\pi(a_{2\ell,m}^2) \sim \text{Gamma}(c_{2\ell}, c'_{2\ell}) \quad 0 \le \ell \le n,$$

we obtain the full conditional distribution for the precision coefficients as

$$\pi(a_{2\ell}^{-2}|\theta_{2\ell,m}(v) : v \in V, -2\ell \le m \le 2\ell) \sim \text{Gamma}\bigg(c_{2\ell} + (2\ell+1/2)|V|,$$
$$c'_{2\ell} + \frac{1}{2}\sum_{m=-2\ell}^{2\ell} \bigg\{ \rho \sum_{v\sim w} \big\{\theta_{2\ell,m}(v) - \theta_{2\ell,m}(w)\big\}^2 + \sum_{v\in V} \theta_{2\ell,m}(v)^2 \bigg\}\bigg).$$

In McMC the angular power spectrum is then updated by sampling independently from these full conditionals and taking the inverse.

5 Results

5.1 Simulation study

We first use simulated data to evaluate the aforementioned method. Synthetic data sets were simulated by randomly selecting a positive tensor under different profiles, the 2nd- and the 4th-order, where we fixed the values of the concentration of water molecules (non-attenuation diffusion) S_0 and the noise variance σ^2. The reference is from real data to resemble the real scenario. We simulated several datasets by choosing different tensor profiles and the noise level σ^2, then compared the performance among

the current most popular methods in DTE, WLS and MLE and the proposed Bayesian method (Bayes). Every dataset contains 1440 measurements which were sampled from 32 distinct gradients (see Table 7) and 15 distinct increasing b-values up to $14000 s/mm^2$ (see Table 5), and the sampling was repeated three times. The ground truth (GT) of high (H-) and low (L-) Rician noise (RN) are 93,0405 and 12,8821, respectively.

We calculated bias between GT and the estimates to evaluate the accuracy and precision of the methods by different criteria: We computed the L^1 norm between GT and $\log \widehat{S_0}$, $\widehat{\sigma^2}$. For the 2nd-order tensor parameter matrix D_0 (GT) and $\widehat{D}$ (estimate), we compared the centered Gaussian displacement distributions $\mathcal{N}(0, D_0)$, $\mathcal{N}(0, \widehat{D})$, with respective densities f_0,$\widehat{f}$, by using the L^2 norm, Kullback-Leibler divergence (KL)

$$\begin{aligned} KL(f_0, \widehat{f}) &= \int_{\mathbb{R}^3} \log\left(\frac{f_0(x)}{\widehat{f}(x)}\right) f_0(x) dx \\ &= \frac{1}{2}\Big\{\log(\det(\widehat{D})) - \log(\det(D_0)) + \ \mathsf{Trace}(\widehat{D}^{-1} D_0) - 3\Big\}, \end{aligned}$$

the symmetric Kullback-Leibler divergence (SKL)

$$SKL(f_0, \widehat{f}) = \big(KL(f_0, \widehat{f}) + KL(\widehat{f}, f_0)\big)/2 \ ,$$

and the Hellinger distance (HL)

$$\begin{aligned} HL(f_0, \widehat{f}) &= \int_{\mathbb{R}^3} \left(\sqrt{f_0(x)} - \sqrt{\widehat{f}(x)}\right)^2 dx \\ &= 2 - 2\det\big((D_0^{-1} + \widehat{D}^{-1})/2\big)^{-1/2} \big(\det(D_0)\det(\widehat{D})\big)^{-1/4}. \end{aligned}$$

Table 1 and 2 illustrate that WLS works well only when the data are less noisy and when the diffusivity is modeled by the 2nd-order tensor. With a truncated dataset, WLS performed better than that with the whole dataset. However, in reality the diffusion MR data are much noisy than the experiment of the low-noise case, and the diffusivity profile is much more complicated than the 2nd-order tensor. Advanced models of diffusion, hence, are needed. Table 2 reveals that when using the whole dataset, WLS is no longer a good choice in comparison to the other alternatives, its estimates of the noise level are strongly biased. MLE under the Rician noise model has overall nice performance as shown in the tables. However, in practice MLE may be very slow to get convergence, and may encounter unstable scenario when the algorithm converges at the local optimum. Our proposed Bayesian approach gives the best performance among the three methods in our experiments, especially in the case with the 4th-order tensor models. These estimators describe the empirical mean values of the posterior rather than the point estimates from maximum a posterior (MAP), see e.g., Andersson (2008) or the other examined methods. Therefore, our method is much more stable regardless and can work with a wide range of data and different diffusivity models. Additional information in Table 1 and 2 includes and units of SKL and HL, they are $\times 10^{-4} mm^2/s$ (6^{**}), and of HL is $\times 10^{-8} mm^2/s$ (15^{**}), respectively, and the number with double asters describes the number of tensor elements that was considered with the metrics.

Table 1. The 2nd-order tensor

L-RN H-RN noise level	$\log S_0$ L1	σ^2 L1	L2	tensor KL	(6**) SKL	(15**) HL
WLS*	0.0011	1.7909	0.1357	3.7778e-05	3.7762e-05	1.8881e-05
	0.0072	7.1720	0.4842	7.7865e-04	7.7317e-04	3.8647e-04
WLS	0.0023	6.8408	0.1570	4.9507e-05	4.9310e-05	2.1298e-04
	0.0068	7.1175	0.3817	4.6489e-04	4.6302e-04	2.4654e-05
MLE	0.0019	0.2132	0.1410	4.0466e-05	4.0299e-05	2.0149e-05
	0.0095	1.2774	0.3864	4.8262e-04	4.8428e-04	2.4210e-04
Bayes	0.0025	0.1272	0.1558	4.9205e-05	4.8991e-05	2.4495e-05
	0.0061	0.8414	0.3387	4.0938e-04	4.0716e-04	2.0355e-04

Unit of SKL is $\times 10^{-4} mm^2/s$, and of HL is $\times 10^{-8} mm^2/s$, and e-04= $\times 10^{-4}$.
$*$ denotes the observations only containing the $b-$values less than $1000 s/mm^2$.
$**$ is the number of tensor elements.

Table 2. The 4th-order tensor

L-RN H-RN noise level	$\log S_0$ L1	σ^2 L1	L2	tensor KL	(6**) SKL	(15**) HL
WLS*	3.3340e-04	0.9630	1.7471	0.0019	0.0018	9.0423e-04
	0.0072	5.5631	2.0119	7.9230e-04	7.8115e-04	3.9043e-04
WLS	8.6455e-04	6.6124	1.4690	0.0018	0.0018	8.8450e-04
	0.0067	47.2370	2.7179	0.0041	0.0038	0.0020
MLE	4.2162e-04	0.1622	1.4143	0.0016	0.0016	8.0334e-04
	0.0095	1.5472	2.2944	0.0012	0.0011	5.6655e-04
Bayes	8.5112e-04	0.2589	0.8443	4.9160e-04	4.8667e-04	2.4328e-04
	6.8639e-04	0.8969	1.2465	5.4696e-04	5.3996e-04	2.6991e-04

Unit of SKL is $\times 10^{-4} mm^2/s$, and of HL is $\times 10^{-8} mm^2/s$, and e-04= $\times 10^{-4}$.
$*$ denotes the observations only contain the $b-$values less than $1000 s/mm^2$.
$**$ is the number of tensor elements.

Computational cost for large-scale data is a common problem in image analysis. Below we illustrate the computational burden (per voxel) in the cases of low and high noise levels, respectively, under the 2nd- and 4th-tensor models by the proposed Bayesian method. In Discussion we will discuss the importance and advantages of Bayesian modeling in DTI.

Table 3. Statistics of McMC convergence

2nd-order		4th-order	
L	H	L	H
nburn/nprec			
7/1832	6/1673	7/1981	7/2010
std			
0.0014	0.0042	0.0015	0.0041
0.1035	0.2843	0.5915	0.9778
0.0417	0.2843	0.6195	1.0340
0.0445	0.1219	0.6130	0.9793
0.1056	0.2781	0.1386	0.2348
0.0450	0.1156	0.1354	0.2298
0.1020	0.2740	0.1197	0.2254
0.3839	2.9403	0.0451	0.1044
		0.0489	0.1198
		0.0367	0.0939
		0.1432	0.3040
		0.1259	0.2759
		0.1692	0.3356
		0.1609	0.3356
		0.1877	0.3556
		0.4079	2.8126

Table 4. List of the computational performance under the 2nd- and 4th-order tensor models

2000 cycles/voxel	2nd-order CT	4th-order CT
CPU time (s) L/H	11.2998/14.9668	33.5773/45.3714

Table 5. b-values (s/mm^2)

62	249	560	996	1556
2240	3049	3982	5040	6222
7529	8960	10516	12196	14000

5.2 Real data

In the follow-up, we illustrate the performance of our method with a real data example.

The dataset The data consists of 4596 diffusion MR-images of the brain of a healthy human volunteer, taken from four $5mm$-thick consecutive axial slices, and measured with a Philips Achieva 3.0 Tesla MR-scanner. The image resolution is 128×128 pixels with size $1.875 \times 1.875\ mm^2$. After masking out the skull and the ventricles, we remain with a region of interest (ROI) V containing 18764 voxels. In the protocol we used all the combinations of the 32 gradient directions listed in Table 7, with the b-values in Table 5, varying in the range $0 - 14000 s/mm^2$, with $2 - 3$ repetitions, for a total of 23 323 644 data points.

McMC implementation The data is analyzed under the 2nd- and 4th-order tensor models, with and without Bayesian regularization, estimating the regularization parameters in the first case. In the Markov chain Monte Carlo we do not impose positivity constraints on the tensors as we discussed in Section 4.3, since we want to count the voxels where the posterior expectation of the tensor is non-positive. To begin with, we compute independently at each voxel v a preliminary estimator for the tensor and noise parameters $\theta(v), \sigma^2(v)$, obtaining the initial state of the Gibbs-Metropolis Markov chain. This is done under the log-Gaussian approximation discussed in Section 2, by the method of weighted least-squares, and using only observations in the low b-value range ($b < 5000\ s/mm^2$). For the regularized model, at each McMC-cycle we divide V into blocks, where each block is the intersection of V with a ball of radius $r = 7$ under the graph distance, and can contain up to 342 voxels. Since blocks are separated by at least one voxel, the parameters from different blocks are conditionally independent given the exterior boundary values, and it is possible to update the blocks in parallel. The centers of the blocks are then cyclically shifted at each McMC cycle, and at the end of each cycle we also update the regularization parameters. The Markov chain was running for 25050 and 22100 cycles respectively, under the 2nd- and 4th-order tensor models, which took 257 and 225 CPU hours on a 15-core Intel Xeon E5-2670 processor.

Monitoring the McMC Before computing empirical averages, we waited for the Markov chain to reach stationarity. The computational time thus depends on how many iterations are needed after burn-in so that the chains get convergence by calling McMC. In Bayesian statistics, we can compute the Monte Carlo (MC) error to determine the number of iterations. Burn-in usually can be dramatically shorten by improving the initials, for example results from MLE or penalized MLE. Our test experiments show that the average burn-in period of one voxel is around 8 cycles and 2000 draws afterwards to get stationarity regardless the tensor models. In this real example, we run 6000 cycles plus 1000 burn-in of all voxels under different tensor models. The computational time at each voxel is dramatically affected by the chosen tensor model as well

as the noise level at that voxel, typically ranging from 10 to 100 seconds. In the updated procedure of McMC, we monitor the log-posteriors in Figure 2, where the traces are from two randomly picked up voxels. The upper indexes indicate the location of the voxels. The chains converge rapidly to stationarity.

The burn-in period used is 1000 cycles for all the selected tensor models. After burn-in, we draw 6000 cycles with the 2nd- and 4th-order tensor models and monitor the logarithmic likelihood, prior and posterior of the samples from two single voxels, and they converged rapidly to stationarity as shown in Figure 1. Such phenomena are not uncommon in high dimensional models, for example the 4th-order tensor models, when MLE is used to construct the initial configuration (see e.g., Figure 3 in Besag et al. (1995)).

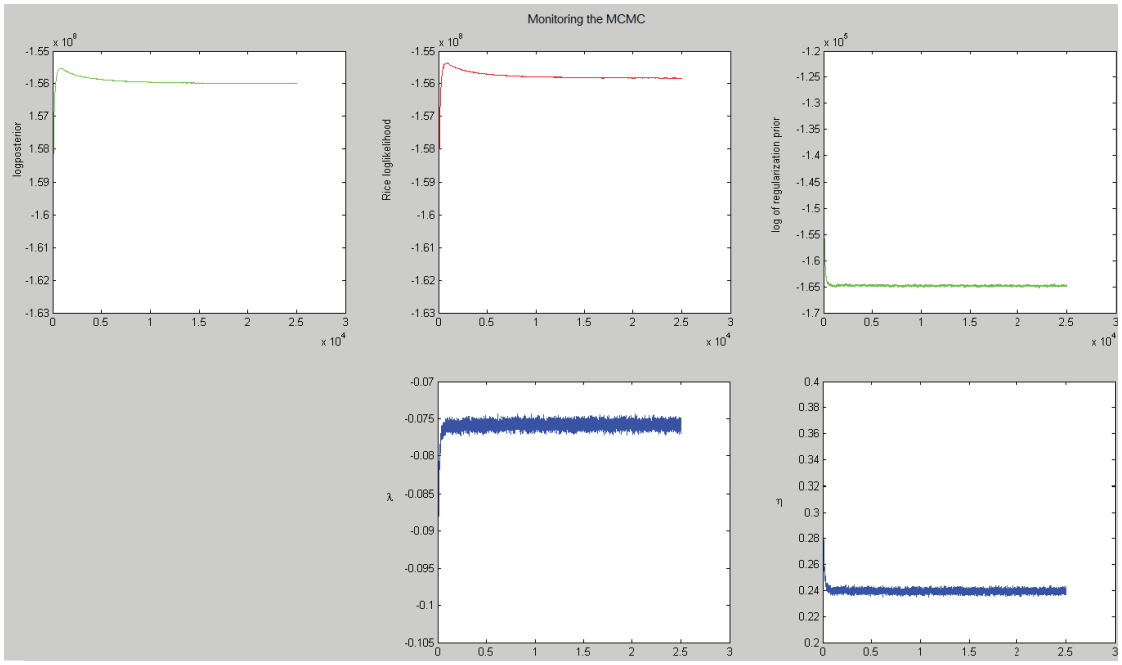
To see this effect in a toy model, just consider a Gaussian vector $X \in \mathbb{R}^n$ with i.i.d. coordinates $X_i \sim \mathcal{N}(\theta, \sigma^2)$, which satisfies

$$\sup_{x \in \mathbb{R}^n} \big\{ \log p_n(x) \big\} - E_P\big(\log p_n(X)\big) = \frac{n}{2} \,. \tag{5.26}$$

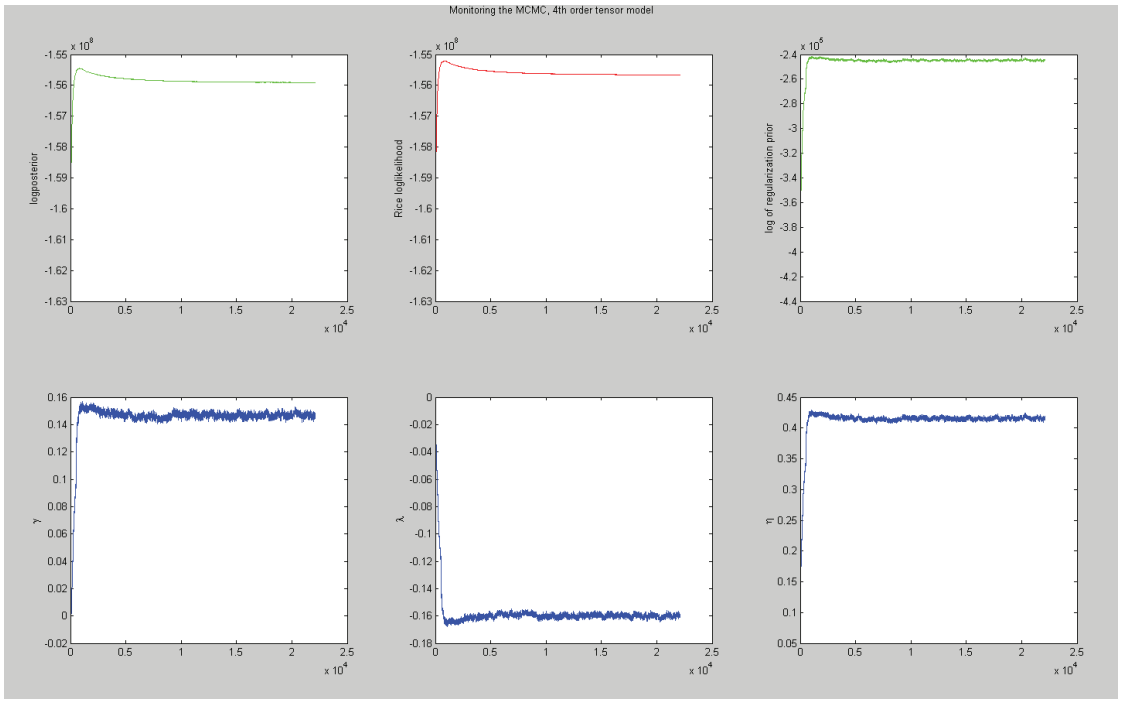
In high dimension, under the posterior distribution the typical configuration and the maximum a posteriori (MAP) configuration can be very different, with a set of typical configurations containing most of the probability mass, while the probability mass concentrated around the MAP-configuration is negligible. Since we start the Markov chain from the maximum likelihood estimator under the approximative log-normal model, at the beginning the orientation of all tensors (but not their eigenvalues) are close to optimal also under the exact Rician likelihood model. Then the tensor eigenvalues and noise parameters move rapidly towards configurations with highest posterior probability. After this phase, it takes a while for the tensor orientations to mix-up. Since the acceptance probabilities are not uniform between blocks and we are estimating simultaneously the regularization parameters, the total log-posterior density shows a slow decay before reaching stationarity.

For comparison, we plot in Figure 2 the McMC trace of the Rician log-likelihood for a single voxel under the 2nd- and 4th-order tensor models, without Bayesian regularization, which converges rapidly to stationarity.

Acceptance probabilities In Figure 3 we show the acceptance probabilities for the Gibbs-Metropolis block update of the tensor parameters, estimated for each voxel under the regularized 2nd- and 4th-order tensor models. Note that, although we use large block updates with more than 300 voxels in each block, the acceptance probabilities are remarkably high in most of the voxels (see the histograms). It means that in most cases our Gaussian approximation is very close to the exact full conditional distribution of the tensor parameters in a block. Note also that in Figure 3a (which corresponds to the 2nd-order tensor model) there are some regions with relatively low acceptance probability. In such areas one should use update blocks of smaller size. These regions of low acceptance probability are either artefacts, where data are corrupted, or contain complex structures where the 2nd-order tensor model does not fit well the data,

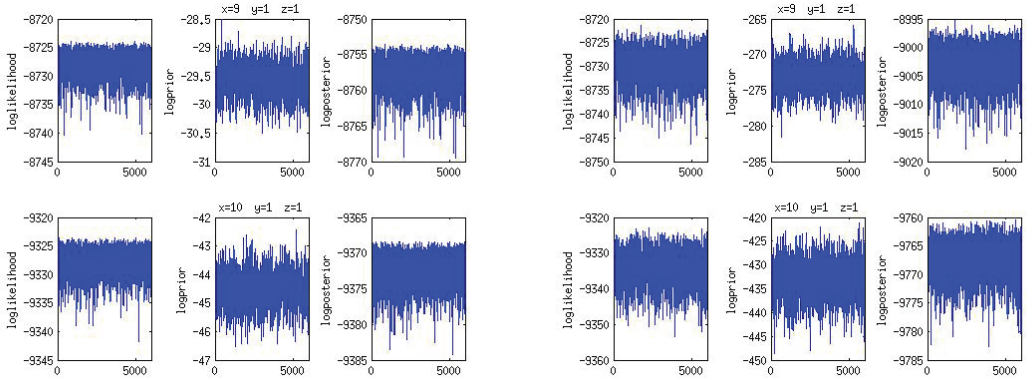


(a) The 2nd-order tensor model, 25050 cycles



(b) The 4th-order tensor model, 22100 cycles

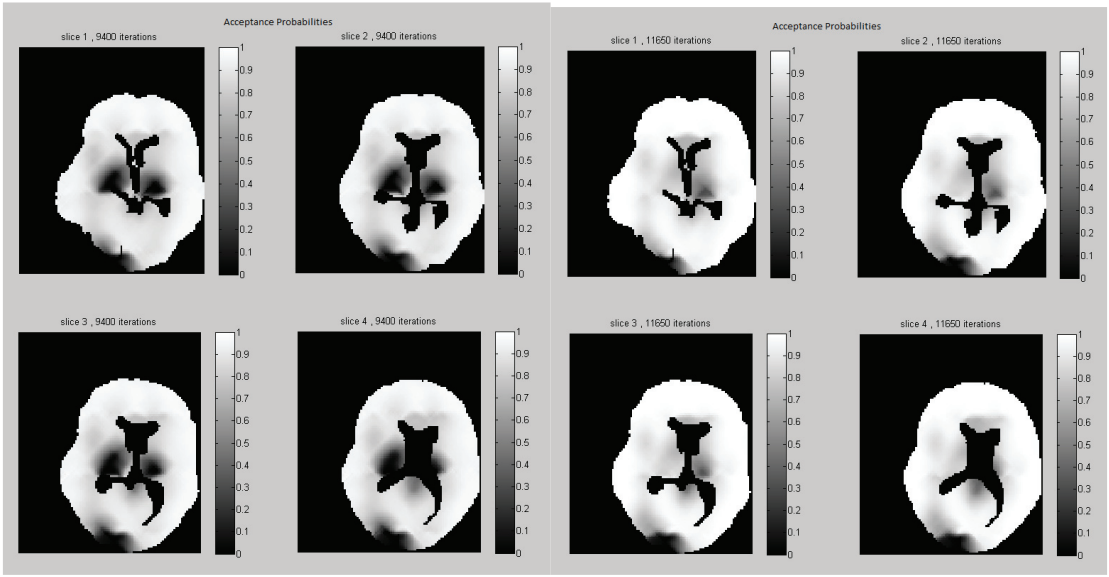
Figure 1. McMC traces of total posterior density, likelihood and prior (in logarithmic scale), and regularization parameters λ, η and γ, for the 2nd- and 4th-order tensor models.



(a) The 2nd-order tensor model, 6000 cycles **(b)** The 4th-order tensor model, 6000 cycles

Figure 2. McMC trace of the Rician log-likelihood for a single voxel, under the 2nd- and 4th-order tensor models (without Bayesian regularization)

and a higher order model would be more appropriate. We see two low acceptance probability regions situated symmetrically on the left and right sides of the ventricles. Anatomically this corresponds to the corona radiata where fiber bundles from multiple directions are crossing. By comparing with Figure 3b we see that in these regions the acceptance probability improves under the (regularized) 4th-order tensor model. For the diffusion model without regularization, the independent tensor updates have high acceptance probabilities at all voxels with both the 2nd- and 4th-order tensor models in Figure 5.



(a) Acceptance probability, the 2nd-order tensor model **(b)** Acceptance probability, the 4th-order tensor model

Figure 3. Acceptance probabilities in gray level scale (black=0,white=1) for the 2nd- and 4th-order regularized tensor models.

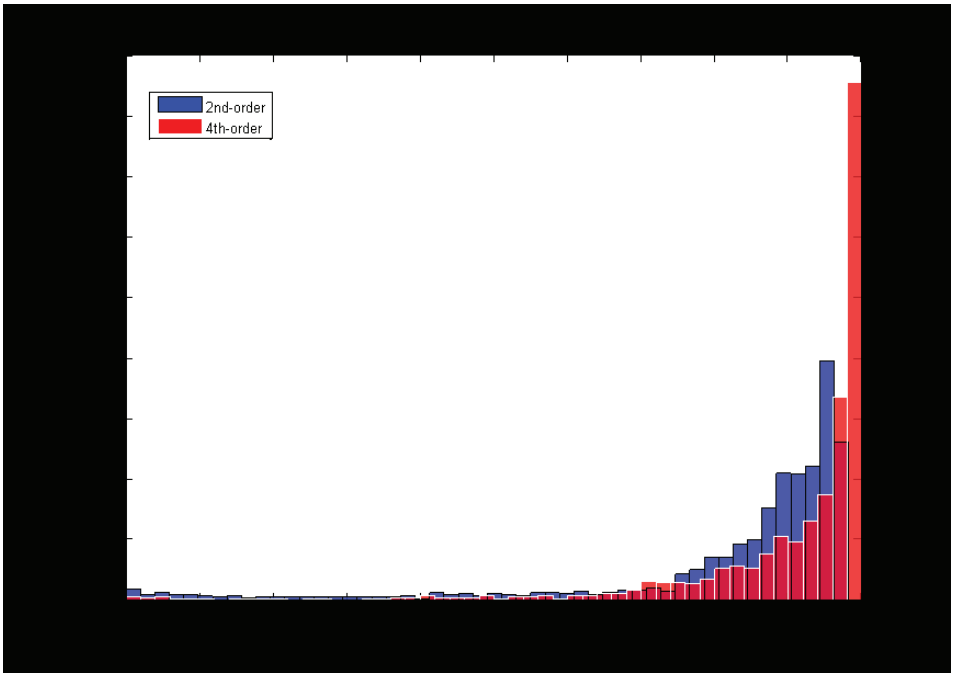


Figure 4. Acceptance probabilities across voxels for tensor block updates, under the 2nd- and 4th-order regularized tensor models.

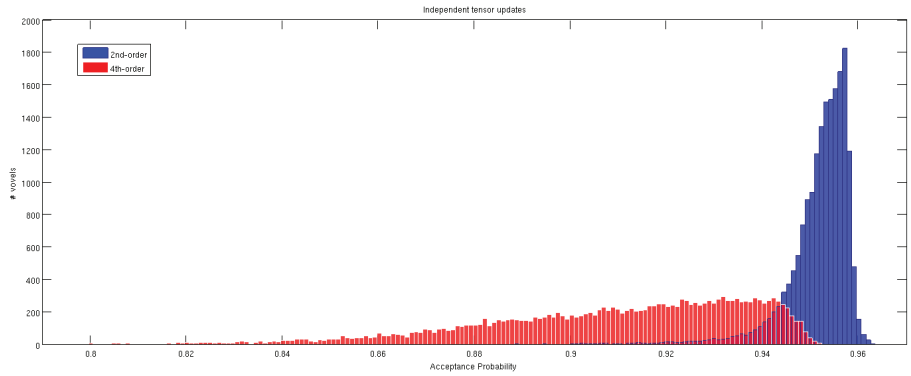


Figure 5. Acceptance probabilities across voxels for tensor independent updates, without regularization, under the 2nd- and 4th-order models.

Deviance Information Criterion The deviance information criterion (DIC), introduced by Spiegelhalter et al. (2002), is a measure of the relative quality of models for given data used in Bayesian model selection as an alternative to Bayes factors. Unlike Bayes factors, DIC is well defined also when improper priors are assumed, as it is the case in our settings. It is defined as

$$\text{DIC} = 2E_\pi\big(D(\theta)\big|\text{data}\big) - D\big(E_\pi(\theta|\text{data})\big),$$

where $D(\theta) = -2\log p(\text{data}|\theta)$ is the deviance, and we take conditional expectations with respect to the posterior distribution of the parameters θ. Defined in analogy with the toy example of Equation (5.26), the *effective number of parameters*

$$n_{eff} := D\big(E_\pi(\theta|\text{data})\big) - E_\pi\big(D(\theta)\big|\text{data}\big)$$

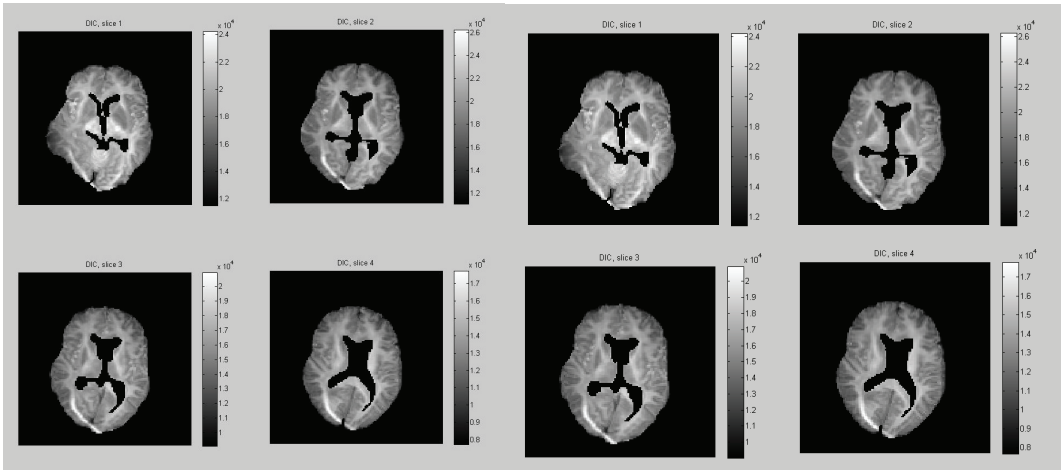
appears as a penalization term in the expression

$$\text{DIC} = -E_\pi\big(\log p(\text{data}|\theta)\big|\text{data}\big) + n_{eff}\ .$$

This allows for model comparisons, lower DIC meaning a better fit to the data relatively to the effective number of parameters. In Figure 6 the DIC is computed independently at each voxel under the 2nd- and 4th-order tensor models (without regularization). Note that the voxels with the highest DIC corresponds to artefacts where data are corrupted, and the area of high DIC correspond to complex white matter structures. We also calculated the overall DIC for all voxels under the 2nd- and 4th-order tensor models with regularization. The respective values DIC$= -1.5554 \times 10^8$ and DIC$= -1.5525 \times 10^8$, indicate that when we penalize the model by the effective number of parameters, overall the 2th-order tensor model fits our data better than the 4th-order model. In Figure 7 the posterior expectation of the noise parameters $\sigma^2(v)$, are shown. When these are interpreted as residual variances in model fitting, we see that they are consistent with the DIC.

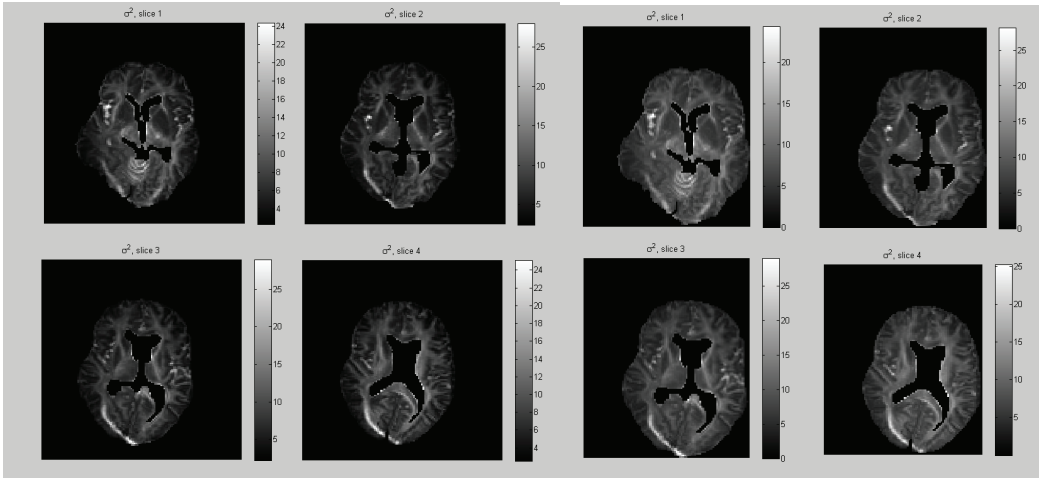
Diffusivity profiles Figure 8 shows the diffusivity profiles based on the posterior estimates of the tensors at all voxels in a region of interest. For each direction $u \in \mathcal{S}^2$ and spatial location $v \in V \subset \mathbb{R}^3$, we plot the point $(v + \overline{d_v(u)}u) \in \mathbb{R}^3$, where $\overline{d_v(u)}$ is the posterior expectation of the diffusivity. In order to observe the differences between the 2nd- and 4th-order tensor models, in Figure 9 we zoom into the ROI (a) and (b), and see that the 4th-order tensor model captures the fiber-crossings which the 2nd-order model cannot capture. At the fiber-crossing locations, under the 2nd-order model the two largest eigenvalues of the estimated tensor have similar sizes, with a donut-shaped diffusivity profile.

Bayesian regularization In Figure 10 we compare diffusivity profiles from a region of interest without and with regularization, under the 4th-order tensor model. With regularization, the differences in shape and direction between neighbouring tensors



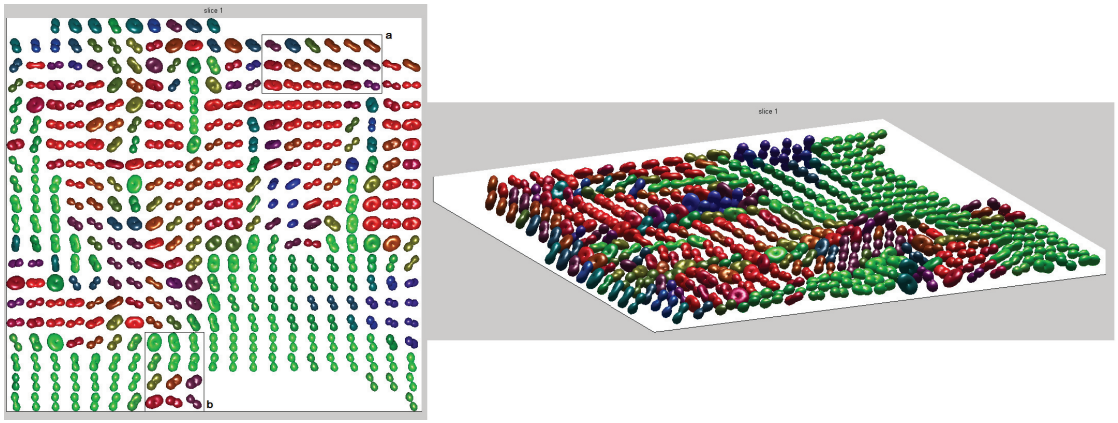
(a) **(b)**

Figure 6. DIC maps under the 2nd- (Figure 6a) and 4th-order (Figure 6b) tensor models, without regularization. Lower (darker) values correspond to better model fit.

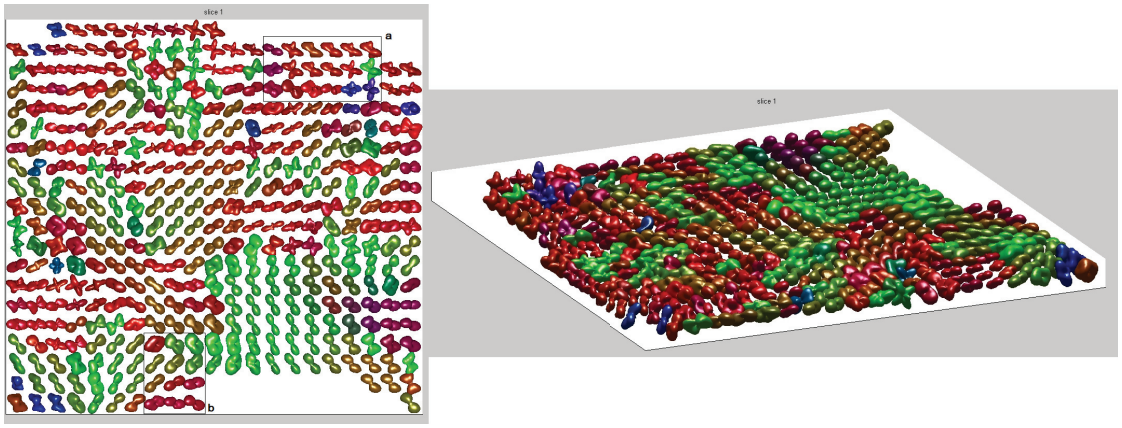


(a) **(b)**

Figure 7. Posterior mean of the noise variance field $\sigma^2(v)$ under the 2nd- (Figure 7a) and 4th-order (Figure 7b) tensor models.

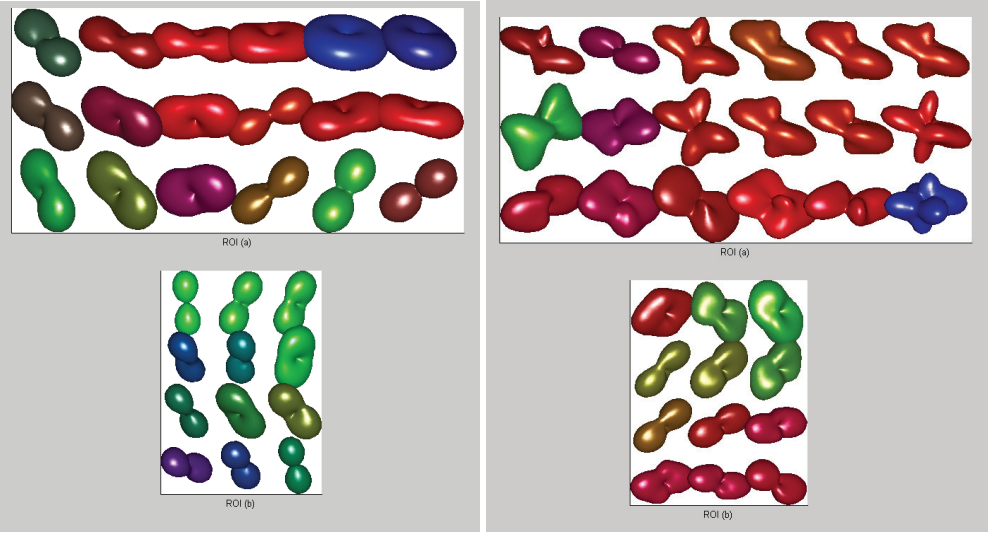


(a) The 2nd-order tensor profile



(b) The 4th-order tensor profile

Figure 8. Estimated diffusivity profiles from a ROI, under the 2nd- and 4th-order tensor model. The color-code represents the main direction of the principal eigenvalue of the 2nd-order tensor: Red, left-right; Green, anterior-posterior; Blue, superior-inferior. These figures are drawn with the MATLAB package FanDTasia written by Barmpoutis A. (Barmpoutis and Vemuri, 2010; Barmpoutis et al., 2009).



(a) The 2nd-order tensor profile **(b)** The 4th-order tensor profile

Figure 9. Estimated diffusivity profiles under the 2nd- and 4th-order tensor models in ROI (a), showing crossing fibers between the corticospinal tract and superior longitudinal fibers, and ROI (b), showing fiber crossing near the corpus callosum, both selected from Figure 9.

Table 6. Posterior mean and standard deviation of regularization parameters under the 2nd- and 4th-order tensor models.

order	$\bar{\eta}$	$\sqrt{\overline{\eta^2} - (\bar{\eta})^2}$	$\bar{\lambda}$	$\sqrt{\overline{\lambda^2} - (\bar{\lambda})^2}$	$\bar{\gamma}$
2	0.2394	0.0012	-0.0758	3.9352×10^{-4}	
4	0.4155	0.0021	-0.16	0.0012	0.1469
order	$\sqrt{\overline{\gamma^2} - (\bar{\gamma})^2}$	$\overline{a_0^{-2}}$	$\overline{a_2^{-2}}$	$\overline{a_4^{-2}}$	
2		0.0029	0.1429		
4	0.0016	0.0029	0.153	1.762	

get smoothed. This also implies noise reduction: the tensor information from data corrupted by artefacts is corrected by the information from the neighbours. For the 2nd-order tensor model, the regularization effect in the same region was not that evident. Since the regularization parameters are not fixed but estimated from data, we cannot always expect an increase from the smoothness level determined by data. In order to achieve a pre-specified level of smoothness we should either fix the regularization parameters or assign them a strongly informative prior. The posterior mean and standard deviation of the regularization parameters are shown in Table 6. The posterior estimates of the inverse angular power spectrum under the the 2nd- and 4th-order tensor models are consistent.

Fractional Anisotropy and Mean Diffusivity. Fractional anisotropy (FA) measures the degree of anisotropy, while mean diffusivity (MD) is the average of the diffusivity

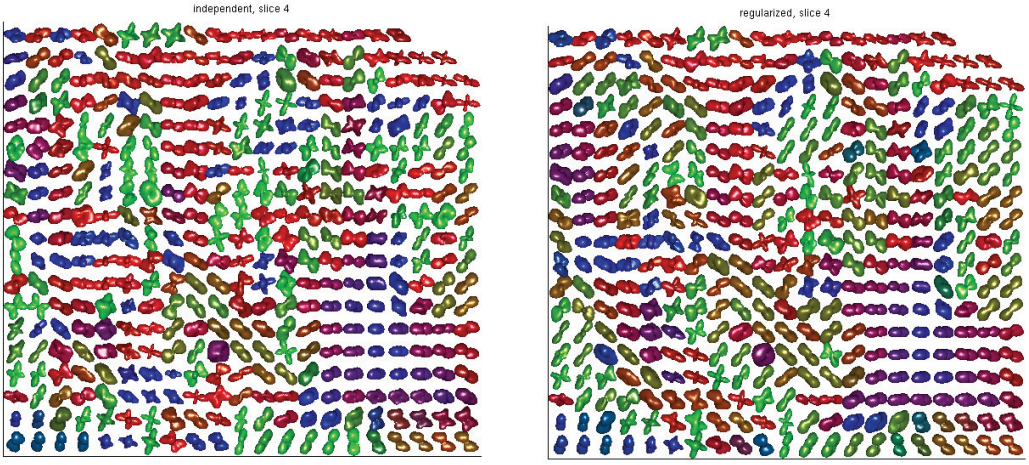


Figure 10. Diffusivity profiles from a ROI under 4th-order tensor model, estimated with and without regularization.

$d(u)$ function over the unit sphere. Both measures are used as biomarkers to study brain pathologies. These quantities are expressed in terms of the eigenvalues of the 2nd-order tensor as

$$\text{MD} = (\lambda_1 + \lambda_2 + \lambda_3)/3, \quad \text{FA} = \frac{\sqrt{3((\lambda_1 - MD)^2 + (\lambda_2 - MD)^2 + (\lambda_3 - MD)^2)}}{\sqrt{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}}.$$

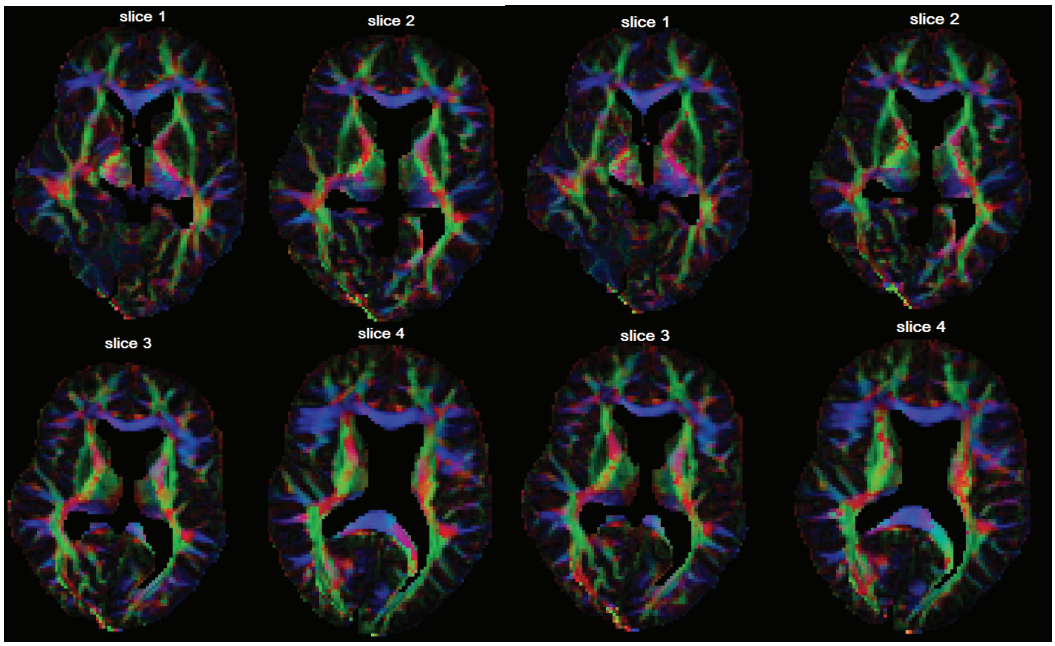
A 4th-order tensor has a map to a 2nd-order tensor as follows according to truncated spherical harmonic expansion of the diffusivity, see details in e.g., Özarslan and Mareci (2003), that is,

$$\begin{aligned}
D_{11} &= \frac{3}{35}(9D_{1111} + 8D_{1122} + 8D_{1133} - D_{2222} - D_{3333} - 2D_{2233}) \\
D_{22} &= \frac{3}{35}(9D_{2222} + 8D_{1122} + 8D_{2233} - D_{1111} - D_{3333} - 2D_{1133}) \\
D_{33} &= \frac{3}{35}(9D_{3333} + 8D_{1133} + 8D_{2233} - D_{1111} - D_{2222} - 2D_{1122}) \\
D_{12} &= \frac{6}{7}(D_{1112} + D_{2223} + D_{1233}) \\
D_{13} &= \frac{6}{7}(D_{1113} + D_{1333} + D_{1223}) \\
D_{23} &= \frac{6}{7}(D_{2223} + D_{2333} + D_{1123}),
\end{aligned}$$

and the mean diffusivity can be also expressed in terms of the 4th-order tensor coefficients as

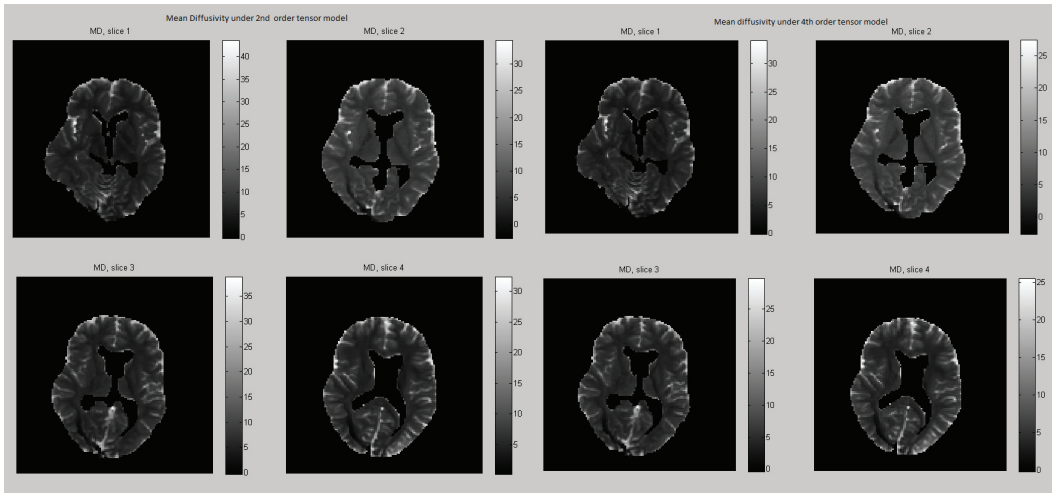
$$\text{MD} = \frac{1}{5}(D_{1111} + D_{1122} + D_{1133} + 2D_{2222} + 2D_{3333} + 2D_{2233} = \frac{1}{5}\text{trace}(\widehat{D}), \qquad (5.27)$$

where $\widehat{D}$ was defined in Equation (4.19). In Figure 11 and 12 we compare the respectively the Bayesian estimates of FA and MD derived under the 2nd- and 4th-order tensor models.



(a) **(b)**

Figure 11. Bayesian FA estimates under the the 2nd-order (Figure 11a) and 4th-order (Figure 11b) tensor models. As in the previous figures, the color-code shows the orientations of the principal eigenvalue of the 2nd-order tensor, with intensities proportional to the fractional anisotropy.



(a) **(b)**

Figure 12. The mean diffusivity (MD) maps from the results for both the 2nd- (Figure 12a) and 4th-order (Figure 12b) diffusion tensor.

6 Discussion

Data augmentation The substantial contribution of this work is the derivation and implementation of a whole statistical strategy of data augmentation. Rician noise, which models the magnitude of a real valued signal perturbed by additive complex Gaussian noise, appears in a wide range of applications in statistics and signal processing. An interesting problem in DTI is to estimate the transition distribution of water molecules diffusing inside the brain cells, by using spectral data which is corrupted by Rician noise. It gives us an opportunity to demonstrate the entire strategy of data augmentation. By using the novel representation of the Rician likelihood, we are able to reduce nonlinear regression problems with Rician noise to generalized linear models with Poissonian noise. Much effort has been devoted to improve the accuracy of the tensor estimation by modeling the noise components appropriately. However, most of early studies are based on a log-normal regression model that assumes the complex Rician noise is additive and has a Gaussian distribution. The model, hence, does not fit the data with high b-values and with low SNR, and does not work with zero measurements.

A fully Bayesian approach A fully Bayesian approach, which provides a way of calculating the full conditional posterior distribution of each parameter of interest, is another contribution of this work. The crucial difference between our method and completing Bayesian methods, such as Andersson (2008), is that we are going to explore the information on the uncertainty of the parameters and further provides better understanding of the estimators by confidence intervals, autocorrelation of marginal covariance, etc. This approach to our knowledge is the best solution in DTI to learn the anatomically or physiologically relevant parameters, such as fractional anisotropy derived from the tensor estimates and the eigenvalues that interpret the fiber directions and take essential roles to probe the tractography of brain connectivity. Furthermore, we introduce a Fisher-scoring algorithm in our Poisson regression model, which gives robust and fast convergence in the MH updates for estimating the tensor parameters. The algorithm hence has helped to shorten the computational execution time and ease the implementation of the Bayesian McMC scheme in our experiments. In addition, the simulation results show that our method provides significantly less biased estimates, especially of the noise variance σ^2 than the alternatives, and the quality of the noise estimates may play crucial role in noise reduction of MRI. Our method also has the best performance among the high angular resolution cases, when compared to the other methods. In conclusion, this Bayesian approach can be considered as a benchmark to evaluate the performance by other methods and of the new derives by qualifying the uncertainty of the parameters.

Model extension and regularization In this work we implement the 2nd- and 4th-order diffusion tensor models. The proposed methods, however, can be easily imple-

mented with most other models of diffusivity in diffusion MRI with proper reparametrization. The Bayesian approach proposed in this work provides an opportunity to model the interactions between voxels or blocks of voxels simultaneously. The results will help us to 1) reduce the noise of DTI based images, and 2) understand the correlations between tensors at voxel level and ROIs in anatomy.

7 Conclusion

Rician noise, which models the magnitude of a real valued signal perturbed by additive complex Gaussian noise, appears in a wide range of applications in statistics and signal processing. By using a novel representation of the Rician likelihood, we are able to reduce nonlinear regression problems with Rician noise to generalized linear models with Poissonian noise. This representation turns out to be very useful in diffusion tensor imaging, where the problem is to estimate the transition distribution of water molecules diffusing inside the brain cells, by using spectral data corrupted by Rician noise. In this work we parametrize these transition distributions with diffusion tensors of either the 2nd- or 4th-order.

Although Bayesian regularization has already been used in the diffusion-MRI literature, until now McMC was not seen as a viable alternative for the analysis of high b-value diffusion-MR data. To obtain diffusion images, we need to process big data. Standard McMC strategies like single site updates and random walk proposals were not efficient enough to produce whole brain images under the Rician noise model. By exploiting the properties of generalized linear models, we are able to construct a Gaussian approximation to the full conditional distribution and update simultaneously large blocks of tensor variables with high acceptance rates. It is clear that our fully Bayesian approach, as well as all methods based on penalized maximum likelihood, is computationally extensive compared with multi-stage procedures where first the tensors are estimated independently, and only in a second step smoothing and interpolation procedures are applied. However second-stage smoothing has its drawbacks, for example it depends on the choice of the tensor metrics, it can induce unwanted effects as tensor swelling (Dryden et al., 2009). Nowadays there are affordable options for acceleration, for example adopting parallel computation on a large computer cluster, and computing with Graphical Processor Unit (GPU) (Hernandez et al., 2013). On the other hand, the acquisition of MR-diffusion data is very costly and we cannot keep a subject for hours inside the scanner, in order to get the most out of the data it makes sense to use more computational resources and perform an accurate Bayesian computation under the true noise model combining estimation and adaptive regularization in a single procedure.

Acknowledgements

We thank Emeritus Professor Antti Penttinen and PhD Salme Kärkkäinen for reviewing the manuscript. The second author was funded by the graduate school of Computations and Mathematical Science (COMAS) of the University of Jyväskylä. We are grateful to the CSC-IT Center for Science Ltd. for the use of their computer cluster, and to the Finnish Doctoral Programme in Stochastics and Statistics (FDPSS) supporting the project with travel grants.

Appendix

A Sampling from the reinforced Poisson distribution

1. The standard way by using the cumulative distribution function:

$$X(\omega) = \min\left\{ n : \sum_{k=0}^{n} \frac{\tau^{2k}}{(k!)^2} \ \geq \ {}_0F_1(1, \tau^2)\, \omega \right\},$$

   with ω uniformly distributed in $[0, 1]$. This requires evaluation of the normalizing constant ${}_0F_1(1, \tau^2)$.

2. A direct but unefficient rejection method:

   Generate $N \sim \text{Poisson}(\tau)$, accept it and set $X = N$ with probability $P_\tau(N' = N|N) = \exp(-\tau)\tau^N/N!$ where N' is an independent copy of N, otherwise repeat until acceptance.

3. An improved rejection sampler, the one actually used. Generate independently $N \sim \text{Poisson}(\alpha)$ and ω uniform in $[0, 1]$,
   until

$$\frac{\tau^{2N}}{(N!)^2} \frac{1}{\pi_\alpha(N)} = \frac{(\tau^2/\alpha)^N}{N!} \exp(\alpha) \geq C(\alpha, \tau)\, \omega$$

   where

$$C(\alpha, \tau) := \max_n \left\{ \exp(\alpha) \frac{(\tau^2/\alpha)^n}{n!} \right\} = \frac{(\tau^2/\alpha)^{n^*}}{n^*!} \exp(\alpha) \tag{A.1}$$

   and $n^* = \lfloor \tau^2/\alpha \rfloor$ is the mode of a Poisson distribution with parameter τ^2/α, and $\lfloor \cdot \rfloor$ denotes the floor function. Return $X = N$.

   For large τ, assuming a priori that at optimality $\alpha \ll \tau^2$, by using Stirling's approximation $\log(n!) \approx (n \log(n) - n)$, we find that the proposal parameter $\alpha(\tau) = \tau$ is approximately optimal.

B Sampling from the log-gamma distribution with small shape parameter.

When the shape parameter is very small, the standard algorithms sampling from a gamma distribution are not reliable. In such cases we use the rejection sampling algorithm for the log-gamma distribution proposed by Liu et al. (2017), described below.

Let X be gamma distributed with shape parameter $0 < a < 1$ and scale parameter 1, then $Y = \log(X)$ is approximated in distribution by $a^{-1}\log(U) = a^{-1}Z$ with U uniform in $[0, 1]$ and $Z = \log(U)$ 1-exponential.

We consider a rejection sampling algorithm for Y with target density

$$p(z) = \Gamma(a)^{-1} \exp\bigl(-z - e^{-z/a}\bigr)$$

and proposal density

$$q(z) = \frac{1}{1+w}\mathbf{1}(z \geq 0)e^{-z} + \frac{w\lambda}{1+w}e^{\lambda z}\mathbf{1}(z < 0)$$

with

$$\lambda = a^{-1} - 1 > 0, \quad w = \frac{a}{(1-a)e} > 0.$$

This is a two sided mixture of exponentials, with parameter 1 on the positive side and λ on the negative side, satisfying the envelope condition

$$\frac{p(z)}{q(z)} \leq \frac{1+w}{\Gamma(a)} .$$

The rejection sampling is implemented as follows: sample independently a proposal value $Z \sim q$ and U uniform in $[0, 1]$, and accept the sample when

$$p(Z)\Gamma(a) \geq \ q(Z)(1+w)U \ ,$$

equivalently

$$\exp(-Z/a) \leq -\log U + (1 - Z/a)\mathbf{1}(Z < 0) \ ,$$

continuing until a proposed value is accepted.

Table 7. For each b-value, the MR-signal was measured in these 32 gradient directions.

u_x	u_y	u_z
-0.5000	-0.5000	-0.7071
-0.5000	-0.5000	0.7071
0.7071	-0.7071	-0.0000
-0.6533	-0.2706	-0.7071
-0.2087	-0.6756	-0.7071
0.0197	-0.7068	-0.7071
0.4212	-0.5679	-0.7071
0.6899	-0.1549	-0.7071
-0.6535	-0.2707	-0.7069
-0.2929	-0.7071	-0.6436
0.2945	-0.7064	-0.6436
0.5150	-0.4861	-0.7061
0.7071	-0.2929	-0.6436
-0.7071	-0.4725	-0.5261
-0.4725	-0.7071	-0.5261
0.5555	-0.6439	-0.5261
0.7071	-0.4725	-0.5261
-0.7071	-0.7071	-0.0002
-0.7071	-0.4725	0.5261
0.7071	-0.4725	0.5261
0.4725	-0.7071	0.5261
-0.7071	-0.7071	0.0078
-0.6364	-0.4252	0.6436
-0.7060	-0.7060	0.0547
-0.2929	-0.7071	0.6436
0.2929	-0.7071	0.6436
0.7071	-0.7071	0.0078
0.7071	-0.2929	0.6436
-0.7063	-0.7063	0.0489
0.0347	-0.7063	0.7071
0.7071	-0.7071	0.0115
0.7071	0.0000	0.7071

References

Andersson, J. L. R. (2008). "Maximum a posteriori estimation of diffusion tensor parameters using a Rician noise model: Why how and but." *Neuroimage*, 42(4): 1340–1356.

Assemlal, H. E., Tschumperle, D., and Brun, L. (2009). "Efficient and robust computation of PDF features from diffusion MR signal." *Medical Image Analysis*, 13(5): 715–729.

Baldeaux, J. and Platen, E. (2013). *Functionals of multidimensional diffusions with applications to finance*, volume 5 of *Bocconi & Springer Series*. Springer, Cham; Bocconi University Press, Milan.

Barmpoutis, A., Hwang, M. S., Howland, D., Forder, J. R., and Vemuri, B. C. (2009). "Regularized positive-definite fourth order tensor field estimation from DW-MRI." *Neuroimage*, 45(1 Suppl): S 153–162.

Barmpoutis, A. and Vemuri, B. C. (2010). "A unified framework for estimating diffusion tensors of any order with symmetric positive-definite constraints." In *Biomedical Imaging: From Nano to Macro. IEEE International Symposium on Biomedical Imaging*, 1385–1388.

Basser, P. J., Mattiello, J., and LeBihan, D. (1994). "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. and Pajevic, S. (2003). "A normal distribution for tensor-valued random variables: applications to diffusion tensor MRI." *IEEE Transactions Medical*, 22(7): 785–794.

— (2007). "Spectral Decomposition of a 4Th-order Covariance Tensor: Applications to Diffusion Tensor MRI." *Signal Processing*, 87(2): 220–236.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). "Bayesian computation and stochastic systems." *Statistical Science*, 10(1): 3–66.

Burdette, J. H., Durden, D. D., Elster, A. D., and Yen, Y. F. (2001). "High b-value diffusion-weighted MRI of normal brain." *Journal of Computer Assisted Tomography*, 25(4): 515–519.

Carr, H. Y. and Purcell, E. M. (1954). "Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments." *Physical Review*, 94(3): 630–638.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging." *The Annals of Applied Statistics*, 3(3): 1102–1123.

Frandsen, J., Hobolth, A., Ostergaard, L., Vestergaard-Poulsen, P., and Vedel Jensen, E. B. (2007). "Bayesian regularization of diffusion tensor images." *Biostatistics*, 8(4): 784–799.

Gasbarra, D., Pajevic, S., and Basser, P. J. (2017). "Eigenvalues of Random Matrices with Isotropic Gaussian Noise and the Design of Diffusion Tensor Imaging Experiments." *SIAM Journal on Imaging Sciences*, 10(3): 1511–1548.

Geman, D., S.and Geman (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6): 721–741.

Ghosh, A., Deriche, R., and Moakher, M. (2009). "Ternary quartic approach for positive 4th order diffusion tensors revisited." In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 618–621.

Ghosh, A., Papadopoulo, T., and Deriche, R. (2012). "Generalized Invariants of a 4th order tensor: Building blocks for new biomarkers in dMRI." In *Computational Diffusion MRI Workshop (CDMRI), MICCAI*, 165–173.

Gradshteyn, I. S. and Ryzhik, I. M. (2015). *Table of integrals, series, and products, edited by Zwillinger, D. and Moll, V.*. Elsevier/Academic Press, Amsterdam, eighth edition.

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

Hahn, E. L. (1950). "Spin Echoes." *Physical Review*, 80: 580–594.

Henkelman, R. M. (1985). "Measurement of signal intensities in the presence of noise in MR images." *Medical Physics*, 12(2): 232–233.

Hernandez, M., Guerrero, G. D., Cecilia, J. M., Garcia, J. M., Inuggi, A., Jbabdi, S., Behrens, T. E., and Sotiropoulos, S. N. (2013). "Accelerating fibre orientation estimation from diffusion weighted magnetic resonance imaging using GPUs." *PLoS ONE*, 8(4): e61892.

Jeffreys, H. (1962). *Cartesian tensors*. Cambridge University Press, New York.

Jones, D. and Basser, P. (2004). "Squashing peanuts and smashing pumpkins: how noise distorts diffusion-weighted MR data." *Magnetic Resonance in Medicine*, 52(5): 979–993.

Koay, C. G., Chang, L. C., Carew, J. D., Pierpaoli, C., and Basser, P. J. (2006). "A unifying theoretical and algorithmic framework for least squares methods of estimation in diffusion tensor imaging." *Journal of Magnetic Resonance*, 182(1): 115–125.

Landman, B., Bazin, P. L., and Prince, J. (2007). "Diffusion Tensor Estimation by Maximizing Rician Likelihood." In *Proceedings of IEEE International Conference on Computer Vision*, 1–8.

Lange, K. (2013). *Optimization, 2nd Edition*, volume 95. Springer.

Lauwers, L., Barbé, K., Van Moer, W., and Pintelon, R. (2010). "Analyzing Rice distributed functional magnetic resonance imaging data: a Bayesian approach." *Measurement Science and Technology*, 21(11): 115804.

Le Bihan, D., Breton, E., Lallemand, D., Grenier, P., Cabanis, E., and Laval-Jeantet, M. (1986). "MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders." *Radiology*, 161(2): 401–407.

Liu, C., Martin, R., and Syring, N. (2017). "Efficient simulation from a gamma distribution with small Shape parameter." *Computational Statistics*, 32(4): 1767–1775.

Liu, J., Gasbarra, D., and Railavo, J. (2016). "Fast estimation of diffusion tensors under Rician noise by the EM algorithm." *Journal of Neuroscience Methods*, 257: 147–158.

Lovász, L. and Vesztergombi, K. (2002). "Geometric Representations of Graphs." In *Paul Erdõs and his Mathematics, (ed. Halász, G., Lovász, L., Simonovits, M. and Sós, V.T.)*, volume 11 of *János Bolyai Society Mathematical Studies*, 471–498. Springer-Verlag.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, 2nd Edition*. Chapman & Hall CRC.

Moakher, M. (2009). "The algebra of fourth-order tensors with application to diffusion MRI." In *Visualization and processing of tensor fields. Laidlaw D.H., Weickert J. (Eds.)*, 57–80. Springer, Berlin.

Mori, S. and Tournier, J. (2014). *Introduction to Diffusion Tensor Imaging 2nd Edition*. Academic Press.

Moseley, M. E., Cohen, Y., Kucharczyk, J., Mintorovitch, J., Asgari, H. S., Wendland, M. F., Tsuruda, J., and Norman, D. (1990). "Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system." *Radiology*, 176(2): 439–445.

Özarslan, E. and Mareci, T. H. (2003). "Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging." *Magnetic Resonance in Medicine*, 50(5): 955–965.

Qi, L., Yu, G., and Wu, E. (2010). "Higher order positive semidefinite diffusion tensor imaging." *SIAM Journal Imaging Sciences*, 3(3): 416–433.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods, 2nd Edition*. Berlin, Heidelberg: Springer-Verlag.

Salvador, R., Pena, A., Menon, D. K., Carpenter, T. A., Pickard, J. D., and Bullmore, E. T. (2005). "Formal characterization and extension of the linearized diffusion tensor model." *Human Brain Mapping*, 24(2): 144–155.

Spiegelhalter, D., Best, N., Carlin, B. P., and van der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 64(4): 583–639.

Stejskal, E. O. and Tanner, J. E. (1965). "Spin Diffusion Measurements: Spin Echoes in the Presence of a Time−Dependent Field Gradient." *The journal of Chemical Physics*, 42(1): 288–292.

Torrey, H. C. (1956). "Bloch Equations with Diffusion Terms." *Physical Review*, 104: 563–565.

Zhu, H., H., Z., G., I. J., and Peterson, B. S. (2007). "Statistical Analysis of Diffusion Tensors in Diffusion-Weighted Magnetic Resonance Imaging Data." *Journal of the American Statistical Association*, 102(480): 1085–1102.

II

FAST ESTIMATION OF DIFFUSION TENSORS UNDER RICIAN NOISE BY THE EM ALGORITHM

by

Liu, J, Gasbarra, D & Railavo, J. 2016

Journal of Neuroscience Methods,
257: 147–158, (2016)

Journal of Neuroscience Methods 257 (2016) 147–158

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

Fast estimation of diffusion tensors under Rician noise by the EM algorithm

Jia Liu [a,*], Dario Gasbarra [b], Juha Railavo [c]

[a] Department of Mathematics and Statistics, University of Jyväskylä, P.O. Box (MaD), FI40014, Finland
[b] Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FI00014, Finland
[c] Helsinki University Hospital, Finland

HIGHLIGHTS

- We originally implement the EM algorithm under data augmentation version in DTI.
- We propose a fast computational scheme for diffusion tensor estimation under the Rician noise model.
- The proposed EM approach is superior in terms of computational burden and estimating accuracy.
- Performance is shown by both mathematical interpretation and numerical comparison.

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form
26 September 2015
Accepted 27 September 2015
Available online 9 October 2015

Keywords:
Data augmentation
Fisher scoring
Maximum likelihood estimator
Maximum a posteriori estimator
Rician Likelihood
Reduced computation

ABSTRACT

Diffusion tensor imaging (DTI) is widely used to characterize, in vivo, the white matter of the central nerve system (CNS). This biological tissue contains much anatomic, structural and orientational information of fibers in human brain. Spectral data from the displacement distribution of water molecules located in the brain tissue are collected by a magnetic resonance scanner and acquired in the Fourier domain. After the Fourier inversion, the noise distribution is Gaussian in both real and imaginary parts and, as a consequence, the recorded magnitude data are corrupted by Rician noise.

Statistical estimation of diffusion leads a non-linear regression problem. In this paper, we present a fast computational method for maximum likelihood estimation (MLE) of diffusivities under the Rician noise model based on the expectation maximization (EM) algorithm. By using data augmentation, we are able to transform a non-linear regression problem into the generalized linear modeling framework, reducing dramatically the computational cost. The Fisher-scoring method is used for achieving fast convergence of the tensor parameter. The new method is implemented and applied using both synthetic and real data in a wide range of b-amplitudes up to 14,000 s/mm^2. Higher accuracy and precision of the Rician estimates are achieved compared with other log-normal based methods. In addition, we extend the maximum likelihood (ML) framework to the maximum a posteriori (MAP) estimation in DTI under the aforementioned scheme by specifying the priors. We will describe how close numerically are the estimators of model parameters obtained through MLE and MAP estimation.

1. Introduction

Diffusion tensor imaging (DTI) is a powerful tool to detect, in vivo, the white matter anatomy and structures of the brain. The raw MR-data are collected by a magnetic resonance scanner and consist of spectral measurement from the displacement distribution of water molecules constrained into cellular structures. Diffusion anisotropy characterizes the nervous fibers.

After the Fourier inversion, the MR-signals are corrupted by a complex Gaussian noise, and consequently, the recorded measurement magnitudes, referred as diffusion weighted magnetic resonance imaging (DW-MRI) data, will follow the Rician distribution. The complex noise is composed of two components, where the real and imaginary parts are still independently Gaussian (Henkelman, 1985; Koay et al., 2009; Zhu et al., 2007). The simplest method for diffusion tensor estimation (DTE) is based on the linearized log-normal regression model, where the residual variance is assumed

* Corresponding author. Tel.: +358 294151407.
E-mail address: jia.liu@jyu.fi (J. Liu).

http://dx.doi.org/10.1016/j.jneumeth.2015.09.029
0165-0270/

to be either constant (the least squares) or depending on the signal amplitude (the weighted least squares). These Gaussian noise models fail to fit the high frequency data, which carry information about the higher order diffusion characteristics. In the existing literature (Rajan et al., 2011; Veraart et al., 2011; Andersson, 2008) on the ML-estimation of diffusion tensors under the Rician noise, the maximization algorithm involves repeated computation of modified Bessel functions. By using data augmentation we are able to replace the Rician likelihood by a Poisson likelihood which is standard in the generalized linear modeling (GLM) framework.

Such simplification reduces dramatically the computational burden of the Fisher-scoring maximization algorithm. This applies also at high b-amplitudes, where in the low signal regime measurements below a threshold are customarily coded as zeros. In the standard LS or WLS approaches, zero-measurements are problematic since they cannot be fitted by a log-normal distribution, and simply discarding them induces selection bias. The appropriately modeled noise level provides capability of data correction in further insights, e.g. removing artefacts from the raw data.

This paper is structured as follows. Section 2 describes the noise in MRI and data augmentation, specifying the statistical model for DTE. In Section 3 we discuss the implementation of the EM and the Fisher-scoring algorithms in the DTI context. In addition, we also specify priors for the parameters and discuss the computation of the maximum a posteriori estimator (MAPE) under the same scheme. Section 4 illustrates the results from both synthetic and real data. Section 5 details the method comparisons. In Section 6 we conclude with an overview of the methods and the undergoing developments. Theoretical details are left for the appendices.

2. GLM for MRI observations

2.1. Rician noise in MRI

In magnetic resonance imaging (MRI), we usually need to take the noise in the raw MR-acquisitions into account. The complex valued noise ε is composed of two *i.i.d.* Gaussian random variables with zero mean and variance σ^2, one for the real and the other one for the imaginary component. After the Fourier inversion, the signal intensity $S \geq 0$ is corrupted by a complex Gaussian noise, and $Y = |S + \varepsilon|$ will be observed. Consequently, the observed MR-signal magnitudes follow a Rician distribution resulting in the likelihood function

$$p_{S,\sigma^2}(y) = \frac{y}{\sigma^2} \exp\left(-\frac{y^2+S^2}{2\sigma^2}\right) I_0\left(\frac{yS}{\sigma^2}\right), \tag{1}$$

where I_α is the α-order modified Bessel function of first kind. For $\alpha = 0$ it has also the following representation in terms of Gaussian hypergeometric series (Jeffrey and Zwillinger, 2007):

$$I_0(2\tau) = {}_0F_1(1, \tau^2) = \sum_{n=0}^{\infty} \frac{\tau^{2n}}{(n!)^2}. \tag{2}$$

Let $t = S^2/(2\sigma^2)$, then Eq. (1) gives

$$P_{t,\sigma^2}(Y \in dy) = \frac{y}{\sigma^2} \exp\left(-t - \frac{y^2}{2\sigma^2}\right) I_0\left(\frac{y}{\sigma}\sqrt{2t}\right) dy \tag{3}$$

with $\tau = yS/(2\sigma^2) = \sqrt{2t}y/(2\sigma)$.

2.2. Data augmentation

We follow the strategy presented in Gasbarra and Liu (2014) implementing augmented data N from a Poisson distribution with mean $t > 0$. The likelihood for the observed data can be transformed from the Rician likelihood equation (3) to a joint augmented density

$$\begin{aligned} P_{t,\sigma^2}(N = n, Y^2 \in dy^2) &= P_{t,\sigma^2}(N = n, X \in dx) \\ &= P_t(N = n) P_{\sigma^2}(X \in dx | N = n) \\ &= \frac{(tx)^n}{(n!)^2 (2\sigma^2)^{n+1}} \exp\left(-t - \frac{x}{2\sigma^2}\right) dx, \end{aligned} \tag{4}$$

where X is from the conditional distribution Gamma$(N+1, 1/(2\sigma^2))$ given N. Eq. (4) provides a transformation from a non-linear regression problem to the GLM framework

$$f_{\xi,\phi}(z) = c(z, \phi) \exp\left(\frac{z\xi - a(\xi)}{\phi}\right) \tag{5}$$

with z corresponding to the response in general, see McCullagh and Nelder (1989) for more details.

3. Method

3.1. DW-MRI and parametrization

In DW-MRI, the signal is modeled as the first equality

$$S(\mathbf{q}) = S_0 \exp(-bd(\mathbf{g})) = S_0 \exp\left(Z\theta\right), \tag{6}$$

where the control vector $\mathbf{q} \in \mathbb{R}^3$ is determined by the sequence of gradient pulses, $b = |\mathbf{q}|^2$, and $\mathbf{g} = \mathbf{q}/|\mathbf{q}| \in \mathcal{S}^2$ is a vector of unit length. The MR-signal decays exponentially with respect to the b-amplitude. Depending on the gradient direction $\mathbf{g}$ the decay is modeled by the reflection symmetric diffusivity function $d : \mathcal{S}^2 \to \mathbb{R}^+$.

Great efforts have been devoted to modeling the diffusivity, and in general we can have parametrization as the second equality in Eq. (6). In the simplest model the diffusivity is expressed by a symmetric and positive definite rank-2 tensor $D \in \mathbb{R}^{3\times 3}$, giving

$$\log S(\mathbf{q}) = \log S_0 - b\mathbf{g}^\top D\mathbf{g} = \log S_0 + Z\theta,$$

where in the left hand side the diffusion tensor is parametrized as

$$\theta = (\theta_1, \ldots, \theta_6)^\top := \left(D_{xx}, D_{yy}, D_{zz}, D_{xy}, D_{xz}, D_{yz}\right)^\top$$

with a design matrix

$$Z = Z(\mathbf{q}) = -b(\mathbf{g}_x^2, \mathbf{g}_y^2, \mathbf{g}_z^2, 2\mathbf{g}_x\mathbf{g}_y, 2\mathbf{g}_x\mathbf{g}_z, 2\mathbf{g}_y\mathbf{g}_z).$$

In high angular resolution models (HARDI) (see, e.g. Barmpoutis et al., 2009), the diffusivity is modeled with a totally symmetric Cartesian tensor D of order $n \in \mathbb{N}$, as

$$d(\mathbf{g}) := \sum_{\ell_1=1}^{3}\sum_{\ell_2=1}^{3} \cdots \sum_{\ell_{2n}=1}^{3} D_{\ell_1,\ell_2,\ldots,\ell_{2n}} g_{\ell_1} g_{\ell_2} \cdots g_{\ell_{2n}}.$$

3.2. EM in MLE

In the optimization of the likelihood, we employ the EM (expectation-maximization) algorithm, which is one among the iterative methods in the MLE or in the maximum a posteriori estimation (MAPE). The EM algorithm proceeds in two steps and shortens the computational complexity by using augmented data. In terms of our case, in the E-step we calculate the expectation of the log-likelihood w.r.t. the conditional distribution of N given by the observations and other parameters with fixed values. In the M-step, we find the ML parameter of S_0^2 and σ^2 by maximizing the augmented log-likelihood quantities. The computational details are listed in Appendix A.

Note that the data are obtained by given different b values and gradients in the experiment, being discrete complex numbers, and therefore, we use sums instead of integrals in the algorithms. The log-likelihood from Eq. (4) is then expressed as

$$Q := \log(p_{t,\sigma^2}(N=n,Y)) = c(Y,N) + N\log(t) - (N+1)\log(\sigma^2) - t - \frac{Y^2}{2\sigma^2}, \quad (7)$$

where $c(Y, N) = N\log(Y^2) - 2\log(N!) - (N+1)\log(2)$ does not depend on (t, σ^2) and will be omitted in the M-step. From Section 3.1, we have $t = S_0^2\exp(2Z\theta)/2\sigma^2$.

In the EM-iteration, given the current parameter estimates $(\theta^{(k)}, S_0^{2(k)}, \sigma^{2(k)})$, we update the conditional expectation of the augmented data by

$$\langle N\rangle^{(k)} := E_{t^{(k)},\sigma^{2(k)}}(N|Y) = \frac{\tau^{(k)}\ I_1\left(2\tau^{(k)}\right)}{I_0\left(2\tau^{(k)}\right)} \quad \text{with}$$

$$\tau^{(k)} = \frac{YS_0^{(k)}\exp(Z\theta^{(k)})}{2\sigma^{2(k)}}.$$

In the M-step we update σ^2 and S_0^2 by the recursions

$$\left(\sigma^{(k+1)}\right)^2 = \left(\sum_{i=1}^{m}\left(\left(S_0^{(k)}\right)^2\exp\left(2Z_i\theta^{(k)}\right) + Y_i^2\right)\right) / \left(2m + 4\sum_{i=1}^{m}\langle N_i\rangle^{(k)}\right) \quad (8)$$

and

$$\left(S_0^{(k+1)}\right)^2 = 2\left(\sigma^{(k)}\right)^2\left(\sum_{i=1}^{m}\langle N_i\rangle^{(k)}\right) / \left(\sum_{i=1}^{m}\left(\exp\left(2Z_i\theta^{(k)}\right)\right)\right), \quad (9)$$

where m is the number of acquisitions at each voxel. For the tensor parameter θ, we employ a stabilized Fisher scoring method: given the stabilizing parameter $\alpha \in [0, 1]$, we iterate the recursion

$$\theta \to \theta + \left((1-\alpha)J(\theta) + \alpha\mathcal{S}(\theta)^\top\mathcal{S}(\theta)\right)^{-1}\mathcal{S}(\theta), \quad (10)$$

until convergence to a fixed point (Lange, 2013). In Eq. (10) the score $\mathcal{S}(\theta)$ is given by

$$\mathcal{S}(\theta) = 2\sum_{i=1}^{m}Z_i\langle N_i\rangle^{(k)} - \left(S_0^{(k)}/\sigma^{(k)}\right)^2\sum_{i=1}^{m}\exp(2Z_i\theta)Z_i^\top,$$

and the corresponding Fisher information is

$$J(\theta) = 2\left(S_0^{(k)}/\sigma^{(k)}\right)^2\sum_{i=1}^{m}\exp(2Z_i\theta)Z_i^\top Z_i.$$

The initials of the EM algorithm can be obtained through the least squares (LS) from a truncated dataset with the diffusion weighting ranging from 0 to 1000 s/mm^2 in order to fit the Gaussian model (see Jones and Basser, 2004; Barber et al., 1998). To pursue higher quality of the initials, we could further apply the weighted least squares (WLS) described in (Zhu et al., 2007). In Appendix B we compare the differences between our EM algorithm and the direct optimization of the Rician likelihood in Eq. (1), which is commonly used to compute the MLE in DTI. It should be noted that the well-known EM algorithm is needed because of the latent augmented variables; it does not decrease the marginal likelihood of the data.

3.3. EM in MAPE

In the Bayesian framework, the maximum a posteriori estimation (MAPE) aims to obtain the point estimates by maximizing the posterior density. The advantage of MAPE over the likelihood approach is that the prior knowledge of the unknown parameters of interest with respect to (w.r.t.) the observed measurements can be transferred into the modeling framework by the prior distribution. Specifically, we can include restrictions to the parameters in terms of probability distributions, for instance regularization can be simultaneously included into the model by adding the knowledge of tuning parameters. Compared with the likelihood approach, Bayesian strategy typically yields less uncertainty and better knowledge of the parameters (the posterior) as it is analyzing the probability distribution of every parameter of interest. The difference between MLE and MAPE in this scenario is in the prior probability $\pi(\xi)$. Given the data y, the normalizing constant in the posterior density $\pi(\xi|y)$ does not depend on the parameter ξ. We find the MAPE by maximizing the joint density $\pi(\xi)p_\xi(y)$, and this is achieved by iterating the EM-recursion

$$\xi^{(k+1)} = \underset{\xi\in\Xi}{\arg\max}\{E_{\xi^{(k)}}(\log p_\xi(z,y)|y) + \log\pi(\xi)\} \quad (11)$$

with the penalization $\log\pi(\xi)$ until convergence to a fixed point. The log-prior penalization term has a regularizing effect, which vanishes asymptotically as the sample size increases (Andersson, 2008).

In DTE, we can assign conjugate priors in light of Section 3.2 for σ^2 and S_0^2. Since we have only weak knowledge of the tensor parameter θ, we may choose non-informative priors which are either scale- or shift-invariant (Jaynes, 2003). A simple Bayesian hierarchical model is obtained after the following choices:

- σ^2 has scale invariant improper prior with density $\pi(\sigma^2) \propto 1/\sigma^2$,
- $S_0^2 \sim$ Gamma(c_1, c_2), where c_1, c_2 are very small.
- $\theta \in R^d$ has the isotropic centered Gaussian prior $\mathcal{N}(0, \Omega^{-1})$, where Ω is a $d \times d$ precision matrix.

The penalized EM-updates for MAPE are given by

$$\left(\sigma^{(k+1)}\right)^2 = \left(\frac{1}{2}\sum_{i=1}^{m}\left(\left(S_0^{(k)}\right)^2\exp\left(2Z_i\theta^{(k)}\right) + Y_i^2\right)\right) / \left(\sum_{i=1}^{m}(2\langle N_i\rangle^{(k)} + 1) + 1\right) \quad (12)$$

and

$$\left(S_0^{(k+1)}\right)^2 = \left(\sum_{i=1}^{m}\langle N_i\rangle^{(k)} + c_1\right) / \left(\frac{1}{\left(\sigma^{(k)}\right)^2}\sum_{i=1}^{m}(\exp(2Z_i\theta^{(k)}) + c_2\right). \quad (13)$$

Additionally, this gives the modified score and Fisher scoring

$$\tilde{\mathcal{S}}(\theta) = \mathcal{S}(\theta) - \Omega\theta \quad \text{and} \quad \tilde{J} = J(\theta) + \Omega, \quad \text{respectively.}$$

Under our Bayesian model with weak priors the MAP estimation equations (12) and (13) are similar as the ML updates Eqs. (8) and (9). Indeed, usually $\sum_{i=1}^{m}\langle N_i\rangle \gg 1$, and we can omit the difference between Eqs. (8) and (12). Then when c_1 and c_2 are small enough, the difference between the likelihood and posterior mode of S_0, expressed in Eqs. (9) and (13) respectively, can also be ignored. The only difference when updating θ is that we have considered the correction between the elements of a tensor represented by the prior distribution, the inverse covariance matrix, Ω. Such a correction may be ignorable.

Remark. By the normalized likelihood, the MLE can be treated as a special case of the MAPE where the precision of the parameters depend on the chosen prior. If the effects of the priors are weak enough to be ignored, then the posterior distribution is asymptotically approximated by the likelihood. The consequence is that numerically the MAP tend to the ML estimates numerically. Such

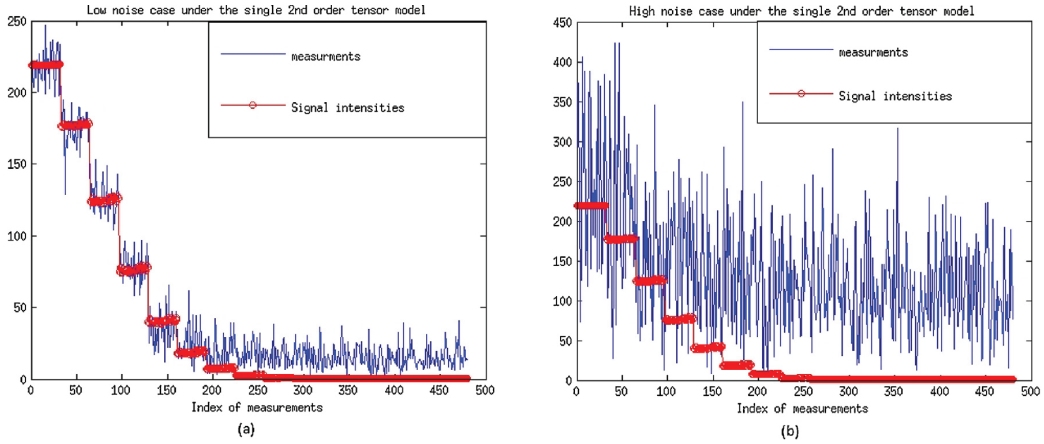


Fig. 1. The thick curve represents the generated data and the red curve gives the corresponding true signal intensities. (a) and (b) describes the generated data and the corresponding true signal intensities under the Rician noise model from the low and the high noise case, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

remark is not unusual (see Sparacino et al., 2000) but nearly has never appeared in the DTI literature.

4. Results

4.1. Synthetic data

4.1.1. Experiment 1

We first simulate four datasets by choosing a positive tensor of 2nd order and of 4th order, respectively from the same voxel with fixed S_0 (5.4595 in logarithmic level) and two different noise variance σ^2. The synthetic data in the experiment arise from models with parameter values (the same gradients, b values and the number of replication which had been used to collect a real human dataset) resembling the real scenario. Each dataset contains 1440 ($32 \times 15 \times 3$) measurements corresponding to 32 distinct gradients and 15 distinct increasing b values (knots), and then being repeated three times. Furthermore, the b knots gradually increase every 32 gradients up to 14,000 s/mm^2 with in total 480 experimental parameters. The ground truth (GT) of high (H-) and low (L-) Rician noise σ are 93.0405 and 12.8821, respectively. Thus we get the (nondiffusion weighted) non-dw SNR (:= S_0/σ) being 2.5256 and 18.2408, respectively, which fall into the wide range of clinic settings (<25) (Veraart et al., 2011). Firstly, we give an overview of the data which are used in this experiment under the signal 2nd order tensor model in Fig. 1. Fig. 1a and b describes the generated data and the corresponding true signal intensities under the Rician noise model from the low and the high noise case, respectively, where we only take the first replication (480 measurements) as an example due to the similar behavior of the other two repeats. From Fig. 1b, we can see that data depicted by the blue curve are much more noisy than that in Fig. 1a. The corresponding diffusion profile of the 2nd order tensor is shown in Fig. 2, where the diffusion profile under the signal 2nd order tensor model represented as an ellipsoid can somehow explain the extent of the departure from normality in the movements of water molecules. In addition, we plot the corresponding diffusion profile of the 4th order tensor in this experiment in Fig. 3, which is also considered to account for possible departures of the observed diffusion from normality.

To compare the performance, we plot the ML estimated signals and the corresponding GT as a function of b values shown in Fig. 4, where we only consider the first 480 measurements as an illustration. The signals are calculated by averaging the 32 gradients for

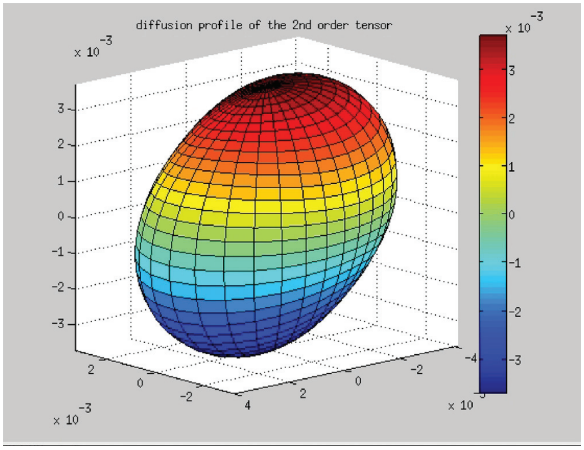


Fig. 2. Scatter plot of the diffusion profile under the selected 2nd order tensor.

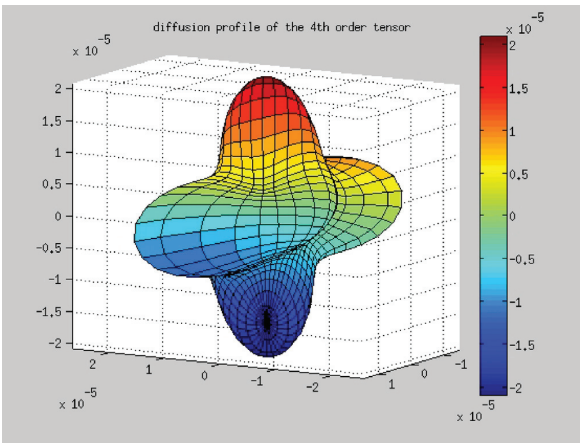


Fig. 3. Scatter plot of the diffusion profile under the selected 4th order tensor.

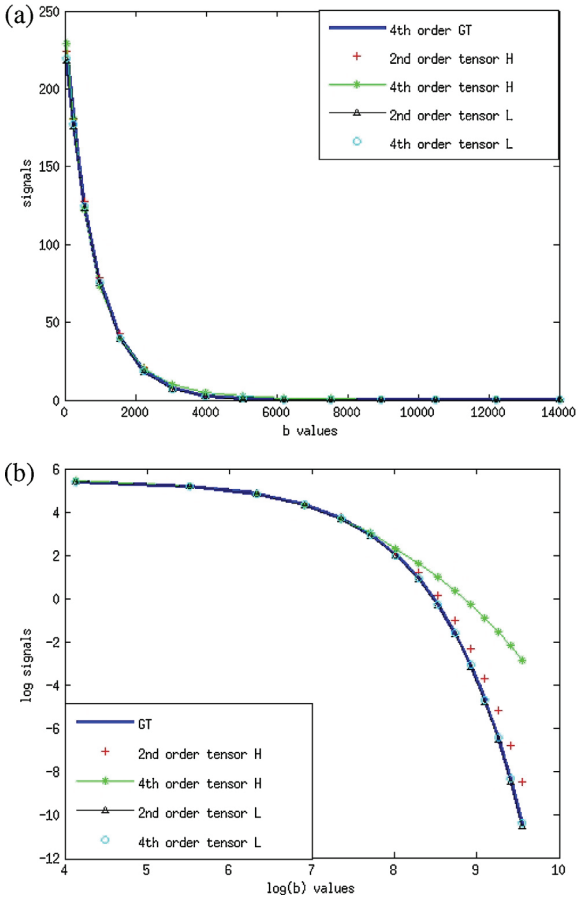


Fig. 4. (a) represents the signals $S(b) = S_0 \exp(Z\theta)$ calculated from the estimated diffusion profile by the proposed MLE method. The thick-blue line depicts the signal intensities of the GT from the 4th order tensor. The green-start line and the cyan circles show the results under the 4th order tensor model from the datasets of the high- and low-noise levels, respectively. The triangular-black line and the red crosses are the results under the single 2nd order tensor model from the datasets of the high- and the low-noise levels, respectively. (b) is the corresponding results in log scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

each distinct b value. In Fig. 4a the signals of ground truth are from the 4th order tensor. However, in reality the ground truth should be unique no matter what choice of angular resolution of the tensor is. Actually in this experiment the signals from the ground truth from 2nd and 4th order have very small difference (the max modulus (m.) deviation in logarithmic scale is less than 0.1, and the mean m. deviation is 0.0374). In order to distinguish the results from different datasets, we plot the results in log scale in Fig. 4b, where we legend the logarithmic signals from the 4th order tensor as GT due to the very small differences mentioned above. The GT are displayed by the thick blue line. As Fig. 4b points out, the results from the dataset under the single 4th order tensor model at the high noise level has 'large' deviation from the GT, but the estimates from the other cases fit the GT quite well. Furthermore, we calculate the empirical signal to noise ratio ($\mathrm{SNR} := (S/SD_v) = (mean(S_g(b,g))/\sigma)$), and only consider one replication. Here instead of averaging the signal intensities of the whole acquisitions as defined in Griffanti et al. (2012), we average the 32 gradients (g) at each distinct b values for representing the changes of the SNR when b value is increasing. To distinguish the difference, we again plot the results from the first 480 measurements in logarithmic level depicted in Fig. 5. It is

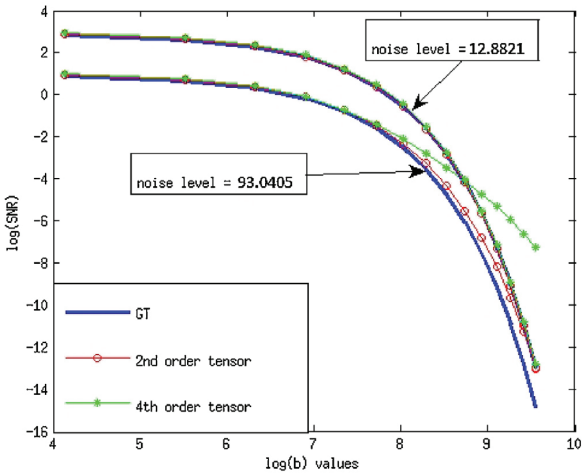


Fig. 5. Empirical logarithmic SNR as a functions of $\log(b)$ values. The GT are represented by the thick-blue lines, of which the upper curve is from the low non-dw SNR corresponding the low noise level with $\sigma = 12.8821$, while the bottom one has the high noise level with $\sigma = 93.0405$. The red-circle lines are the fitted profile under the single 2nd order tensor model, and the green-star lines show the empirical SNR under the single 4th order tensor model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

shown that in the high-noise level case, the results under the single 4th order tensor model have a bit larger bias when $b \geq 3000$ s/mm^2.

4.1.2. Experiment 2

For comparison of the methods, we generate 100 datasets from the high (Figs. 6–8) noise case and another 100 datasets from the low (Figs. 9 and 10) noise case under the same 4th order tensor as in Experiment 1 and compare the sample means of SNR ($\mathrm{SNR} := (S/SD_v) = (mean(S_{g,r}(b,g,r))/\sigma)$) of the whole 1440 measurements in each sample data with the corresponding GT from the different methods, where the mean of the signals in the numerator is calculated by averaging the 32 gradients (g) and the total

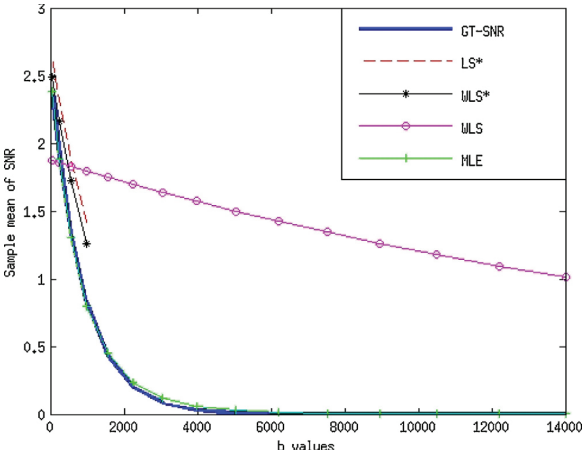


Fig. 6. Sample mean of SNR as a function of b values. The sample means are calculated from 100 simulated datasets. The SNR are calculated from the estimates estimated by the different methods. The thick-blue curve represents the SNR of the GT. The red-dash line and the black-star line are the estimators by the LS and the WLS with the truncated datasets, respectively. The cyan-circle line is the results through the WLS, and the green-cross line is empirical values by our MLE method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

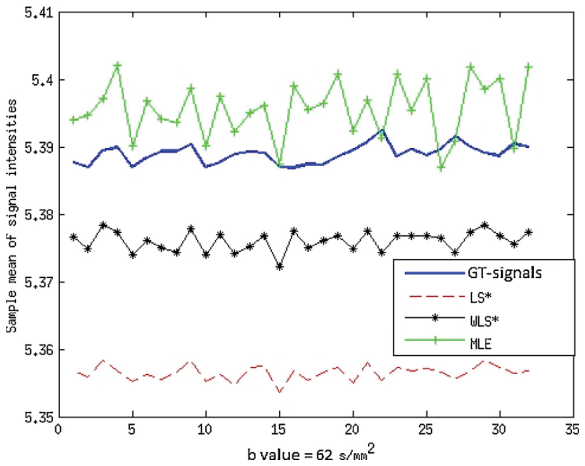


Fig. 7. Sample mean of signal intensities. Again the thick-blue curve represents the GT. The red-dash line and the black-star line are the results by the LS and the WLS methods with the truncated datasets, respectively. The green-cross line shows the results by our MLE method. We did not show the results by the WLS from the whole dataset as the bad performance in Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

number of replications (r) from the whole measurements. Note that here we also average the number of replication in each dataset. Fig. 6 represents the results from the datasets generated by the high-noise level, where "*" denotes that only the low frequencies (b values less than 1000 s/mm^2) are considered in the estimation. This figure reveals that the fitting profile by our method is the best, while the WLS results from the whole data space are much worse than the others. To compare the further performance, we compute the sample mean of signal intensities, and as an example, we pick up from the first replication those intensities with a low b value. The result is in Fig. 7, from which, we can see that our results are slightly over-estimated from the high-noise level data, but still being the best. The results from the other two methods are under-estimated. In addition, we compute the sample mean

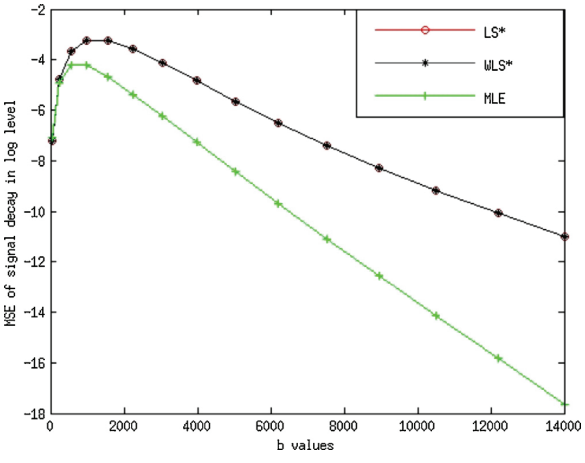
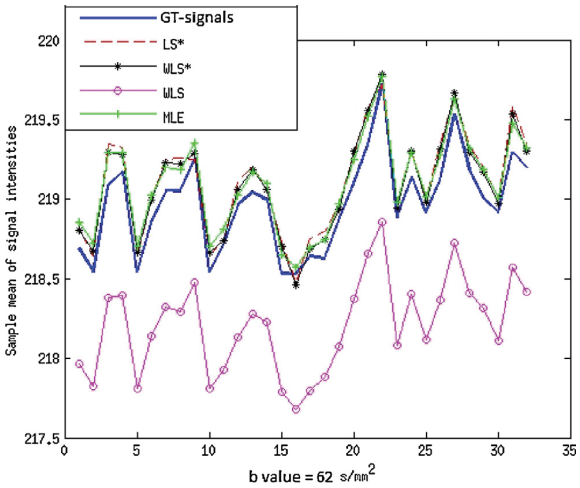
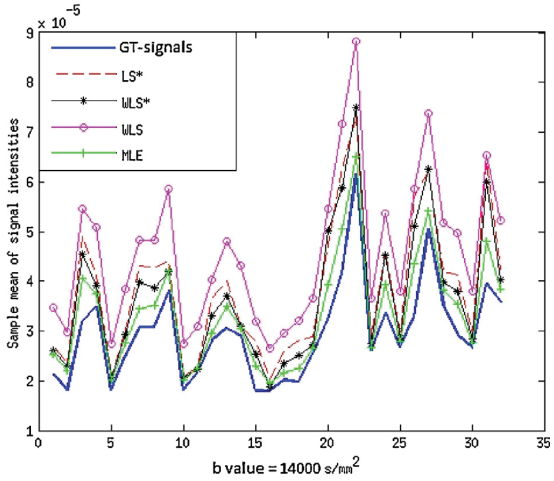


Fig. 8. MSE of sample mean of averaged signal decay as a function of the distinct b values from the first 480 measurement. The red-circle line and the black-star line the results by the LS and the WLS methods with the truncated datasets ($b \leq 1000$ s/mm^2), respectively. They are almost overlapping. The green-cross line shows the results by our MLE method. We did not show the WLS results from the whole dataset due to the bad performance in Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) low b value, $b = 62 s/mm^2$



(b) high b value, $b = 14000 s/mm^2$

Fig. 9. Sample mean of signals intensities. The plots illustrate the means of signal intensities at $b = 62$ and 14,000 s/mm^2, respectively, of each gradient from the first replication estimated by the four methods. The red-dash line and the black-star line are the results by the LS and the WLS methods with the truncated datasets, respectively. The green-cross line show the results by our MLE method, and cyan-circle line is the results through the WLS. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of signal decay $S(b)/S(0) := \exp(Z\theta)$ by the tensor coefficients averaging the gradients for obtaining the mean square errors. Fig. 8 describes the mean square error of signal decay in log level as a function of b values. Note that the results by the LS* and the WLS* are extrapolated to the high-frequency region by using the same design matrix Z and their tensor estimates. This figure reveals that even in the region of low b values ($b = 800$–1000 s/mm^2), our method still performs better than the others.

Figs. 9 and 10 correspond with Figs. 7 and 8 from the 100 sample data generated by the low noise. Fig. 9a reveals that the estimated signal intensities from our method are roughly similar than the results from the LS* and the WLS* when the b value equals to 62 s/mm^2. In Fig. 9b again the signal intensities by the LS* and

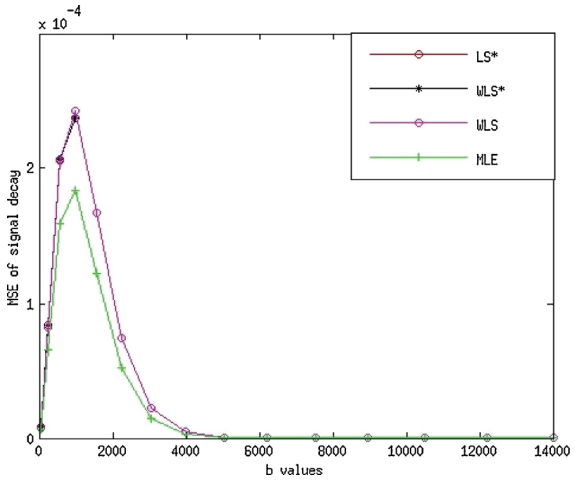


Fig. 10. MSE of sample mean of averaged signal decay as a function of the distinct b values from the first 480 measurement. The red-circle line and the black-star line the results by the LS and the WLS methods with the truncated datasets ($b \leq 1000$ s/mm^2), respectively. They are almost overlapping with the results by the WLS. The green-cross line show the results by our MLE method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the WLS* are extrapolated to the high-frequency region by the estimated diffusion profile, and our method shows a better fitted profile than the others. Fig. 10 describes the mean square error of signal decay as a function of b values. Since the difference is visible, we do not need rescale the results in the log level. This figure reveals at $b \leq 1000$ s/mm^2, the LS*, WLS* and the WLS methods perform quite similarly, and the results by our method represent the smallest MSE in the whole region of the frequencies.

All the synthetic experiments were carried out on a 64-bit 4 core computer with 16 Gb RAM, and the CPU of each core is 3.40 GHz with MATLAB. The average computational time of the aforementioned MLE method under the 4th order tensor model is 0.5435 s (an example record from the 100 datasets under low noise case), which is extremely shorter than the minutes running time per voxel from the current standard methods such as MATLAB Nelder–Mead based or gradient-based estimators (see Ghosh et al., 2014; Landman et al., 2007).

4.2. Real data

The data consist of 4596 diffusion MR-images of the brain of an healthy human volunteer, taken from four 5 mm-thick consecutive axial slices, and measured using a Philips Achieva 3.0 Tesla MR-scanner. The image resolution is 128×128 pixels of size 1.875×1.875 mm^2. After masking out the skull and the ventricles, we remain with a region of interest (ROI) containing 18,764 voxels. In the protocol, we used all the combinations of the 32 gradient directions with the b-values varying periodically in the range 0–14,000 s/mm^2, with 2–3 repetitions, for a total of 23,323,644 data points. The average computational cost per voxel by our method under the 4th order tensor model from this dataset is 1.8331 s. We illustrate the results mainly under the 4th order tensor model. Fig. 11 shows the mean diffusivity (MD) and the fractional anisotropy (FA) of diffusion from two consecutive slices, where FA is computed from the results under the 2nd order tensor model, given by

$$\mathrm{FA} = \frac{\sqrt{3((\lambda_1 - E[\lambda])^2 + (\lambda_2 - E[\lambda])^2 + (\lambda_3 - E[\lambda])^2)}}{\sqrt{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}}. \tag{14}$$

The average values of FA from these two ROI are 0.2769 and 0.2861, respectively. The color in FA represents the orientations of the fibers. Under the 4th order tensor model, MD is expressed as

$$\begin{aligned}\mathrm{MD} &= \frac{1}{5}(D_{1111} + D_{2222} + D_{3333} + 2D_{1122} + 2D_{1133} + 2D_{2233}) \\ &= \frac{1}{5}\mathrm{trace}(D).\end{aligned} \tag{15}$$

The average values of MD from Slice 3 and 4 are 6.248e−03 mm^2/s, 6.045e−03 mm^2/s, respectively, and we have the same estimated values of MD under the 2nd order tensor model.

We also plot the Rician noise map of σ from the two consecutive slices shown in Fig. 12, where the artefacts are clearly depicted by white color representing very high noise, which reveal the true scenario from the raw MR images, and are confirmed independently by our estimation.

Visualization of angular resolution of DTI data under different tensor models from the region of interest (ROI) of two consecutive slices are displayed in Fig. 13, where the ROI is near the hippocampus and the empty spaces inside of left parts of the diffusion profiles (DP) are the masked ventricle. DP under the 4th order tensors provide detailed information of diffusion through the higher angular resolution. In addition, the colors represent the principle orientations of diffusion at each voxel. These tensor profiles are plotted by MATLAB fanDTasia toolbox (Barmpoutis et al., 2007). We also conduct the experiment with the real data on the 64-bit 4 core computer with 16 Gb RAM, and the CPU of each core is 3.40 GHz with MATLAB. The total running time is 2.9733e+04 and 3.4395e+04, equally 1.5846 and 1.8331 s per voxel in average under the 2nd and 4th order tensor model, respectively.

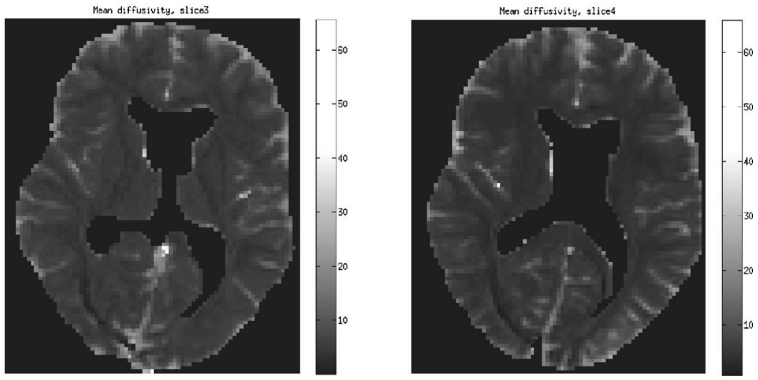
Note that the algorithms presented in this work are under the assumption of voxel independence, therefore, the algorithms are parallelizable across voxels. The code related to the proposed method and the above results is available by request, which can also work on the cluster by parallel computation pixel by pixel.

5. Method comparisons

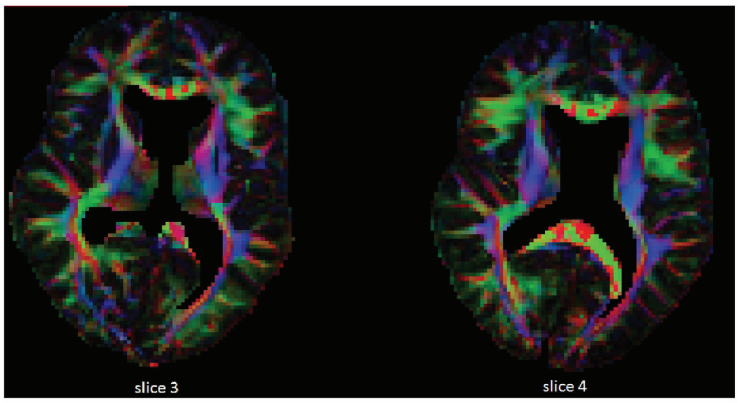
5.1. Comparison between our EM method and the traditional MLE (Andersson, 2008)

In this section, we discuss the differences between our data-augmentation based on the EM algorithm and on the typical MLE method through direct maximization at the Rician log-likelihood Q_r. Detailed calculation can be found in Appendix B.

1. We do not need to calculate all the elements of the Hessian as we can directly find the modes of S_0^2 and σ^2 by data augmentation. A small improvement appears in the reparametrization of S_0 or $\log S_0$ by S_0^2.
2. In the E-step we compute
$$\langle N_i \rangle = E_{\theta^{(k)},\sigma^{2(k)},S_0^{2(k)}}(N_i|Y_i), \tag{16}$$
which does not depend on the parameters θ, σ^2 and S_0^2. In the M-step we use Eq. (16), the recursive values from $\theta^{(k)}$, $\sigma^{2(k)}$, $S_0^{2(k)}$, instead of solving the intractable formula w.r.t. those parameters. This dramatically reduces the computation of the score from Eqs. (B.2)–(B.3)) to Eqs. (A.3)–(A.4)), respectively.
3. The EM algorithm allows us to use empirical values from Eq. (16) to compute the Fisher information. Our Fisher information $J(\theta)$ which fits the whole range of SNR and is slightly bigger than the approximated one, $\mathcal{I}_r(\theta)$, expressed in (Eq. (B.4)), which requires heavy mathematical calculations to deal with different expectations (see Andersson, 2008 for more details). In addition, when computing the score of θ in Eq. (10), we do not need to update



(a) Mean diffusivity (MD)



(b) Fractional anisotropy (FA)

Fig. 11. MD and FA maps from two consecutive slices, where the estimated FA are computed under the 2nd order tensor model. The color in FA represents the orientations of the fibers: red, left–right; green, anterior–posterior; blue, superior–inferior. The color coded FA maps are drawn by using the software ExploreDTI (Leemans et al., 2009). The corresponding MD maps are from the results under the 4th order tensor model, where the white spots corresponding to the corrupted data (artefacts) with measured magnitudes increasing to high *b* values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

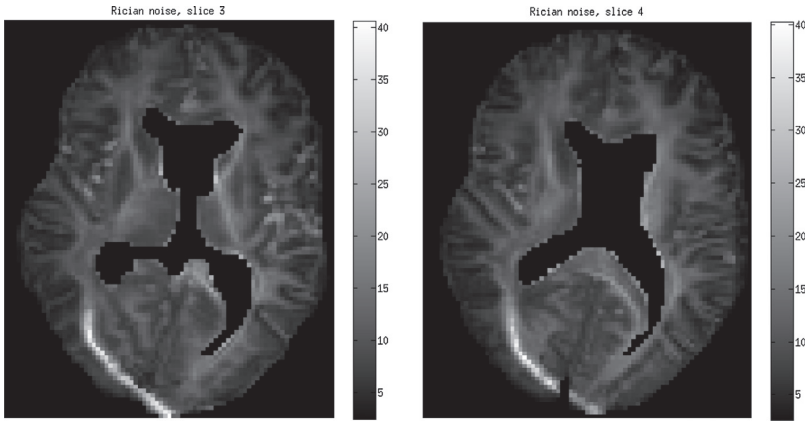


Fig. 12. Rician noise map from two consecutive slices. The white curves in the left bottom of the slices depict the artefacts corresponding to very high noise.

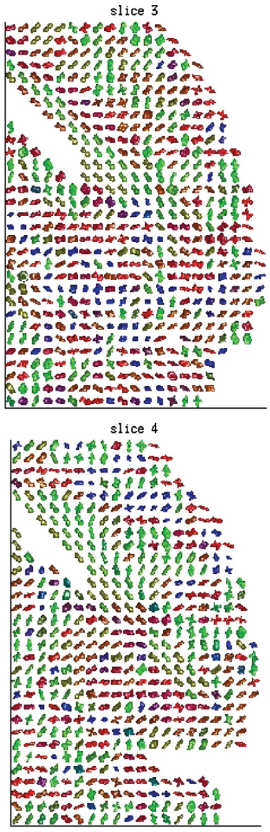


Fig. 13. Visualization of the 4th order diffusion tensor profiles from two consecutive slices of a ROI. The color-code represents the main principal direction of the diffusion: red, left–right; green, anterior–posterior; blue, superior–inferior. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the items containing N_i as they are fixed values from Eq. (16). All those lead to reduced computation in practice.

5.2. Comparison between our EM method and the EM method described, e.g. in Solo and Noh (2007) and Zhu et al. (2013)

Firstly, the theory part of the comparing EM method can be found in Appendix C.

1. In terms of the EM algorithm, both methods are likely in computation. Since the augmented data are calculated in the E-step by the knowns and parameters in the current iteration given by, respectively

$$\langle N\rangle^{(k)} := E_{S^{(k)},\sigma^{2(k)}}(N|Y) = \frac{\tau^{(k)}\ I_1\left(2\tau^{(k)}\right)}{I_0\left(2\tau^{(k)}\right)},$$

$$\langle\cos(\varphi)\rangle^{(k)} = E_{S^{(k)},\sigma^{2(k)}}(\cos(\varphi)|Y) = \frac{I_1(2\tau^{(k)})}{I_0(2\tau^{(k)})} \quad \text{with}$$

$$\tau^{(k)} = \frac{YS_0^{(k)}\exp(Z\theta^{(k)})}{2\sigma^{2(k)}}.$$

In the M-step, we calculate the partial derivative of Q w.r.t. σ^2 and S_0. Such derivatives are straightforward to compute as presented in Zhu et al. (2013). Then the computation till now from both methods should be roughly similar. The difference is that, in our EM algorithm, we update θ, the tensor parameter by a stabilized Fisher scoring method.
2. In theory, the augmentation in the two EM algorithms have essential difference, that is, they are working in different space. The implemented augmentation is in the natural integer space, while the introduced augmentation in Appendix C works on the phase data space.
3. In terms of Bayesian strategy, both methods can be totally different, because we can include the prior knowledge of the argument data through the prior distributions, then N will be generated from the reinforced Poisson distribution (see (Gasbarra and Liu, 2014)) and $\cos(\phi)$ will be obtained from the Von Mises distribution given in Eq. (C.1).

6. Discussion

Our method substantially differs from the previous ones in the literature and the advantages are summarized by the following points: (1) We implement the recently developed data augmentation method (Gasbarra and Liu, 2014), which allows the non-linear regression problem to be transformed into the GLM framework in DTE. (2) Subsequently, the computation is dramatically reduced due to the tractable modes of parameters of interest in the sense of point estimation. In addition, when employing Fisher-scoring scheme we simplify the complexity of the Fisher information. (3) Our Rician noise model can be combined with any tensor model in different representation, such as spheric harmonic expansion, by reparametrization. (4) Either ML or MAP estimation yields more accurate estimates than the LS and the WLS do. In addition, high frequencies from the low SNR data and the zero measurements are also included into the estimation. These data are known to contain detailed anatomical information of the complex tissue in vivo. (5) Our method leads to significantly less biased estimates of the noise level, which plays a key role in denoising the MRI and cleaning the artefacts.

Positivity constraints. The physical feature of diffusion requires the tensor to be positive definite. Our model allows to check the positivity of diffusivity in the tensor updates under the scheme of Fisher-scoring method. For the rank-2 tensor model, the constraining is fairly easy to do by computing the eigenvalues of the tensor matrix D. For HARDI, Barmpoutis et al. (2009) propose the Gram matrix approach, using the quartic form to guarantee the positivity. Other methods such as Qi et al. (2010) address the constraint by calculating the Z-eigenvalue polynomials.

MLE vs. MAPE. In this work, we did not list the results from MAPE but we emphasize the differences between these two methods. Bayesian methods have advantages in the learning process, meaning that they may gain extra information from the prior knowledge. When the prior is weak, like in our case, we learn things from the data, what we actually do when approaching the problem through frequentist statistical modeling. In order to learn the uncertainty of the diffusion parameters, a fully Bayesian approach is highly recommended to characterize the posterior parameter distributions rather than point estimation.

Acknowledgements

First we would like to thank the two anonymous reviewers for the invaluable comments which have led to the significant improvement of this manuscript. We thank Professor Antti Penttinen for carefully reading the manuscript and providing insightful comments. We would also like to thank the Radiology Unit of Helsinki University Hospital and

Professor Oili Salonen for supporting the data collection. This work was funded by Doctoral Program in Computing and Mathematical Sciences (COMAS), University of Jyväskylä. We acknowledge the Finnish Doctoral Programme in Stochastic and Statistics (FDPSS) provided travel funds for this research.

Appendix A. MLE by the EM algorithm in DTI

We consider the Rician noise model with the Poissonian data augmentation of Section 2. The latent augmented variable N conditionally on X, Z is given by

$$p_{t,\sigma}(N=n|X,Z)=\frac{1}{I_0(2\tau)}\frac{\exp(-2\tau)\tau^{2n}}{(n!)^2},$$

$$n\in\mathbb{N}\quad\text{with}\quad\tau=\sqrt{\frac{Xt}{2\sigma^2}}\quad\text{and}\quad X=Y^2.$$

It follows Gasbarra and Liu (2014) that this discrete distribution is referred as reinforced Poisson distribution with parameter τ.

In the EM algorithm we need to compute the conditional expectation of N conditionally on X and the design matrix Z. Given the current values $t^{(k)}, \sigma^{2(k)}$, then

$$\begin{aligned}\langle N\rangle^{(k)} :&= E_{t^{(k)},\sigma^{2(k)}}(N|X,Z)=\sum_{n=1}^{\infty}np_{t,\sigma}(N=n|X,Z)\\&=\tau^{(k)}/2\frac{d}{d\tau^{(k)}}\log{}_0F_1(1,(\tau^{(k)})^2)\\&=\tau^{(k)}/2\frac{d}{d\tau^{(k)}}\log J_0(2\tau^{(k)}\sqrt{-1})\\&=\frac{\tau^{(k)}J_{-1}(2\tau^{(k)}\sqrt{-1})}{J_0(2\tau^{(k)}\sqrt{-1})}=\frac{\tau^{(k)}\ I_1(2\tau^{(k)})}{I_0(2\tau^{(k)})},\end{aligned}$$

with

$$t^{(k)}=t(S_0^{2(k)},\theta^{(k)},\sigma^{2(k)})=\frac{S_0^{2(k)}\exp(2Z\theta^{(k)})}{2\sigma^{2(k)}},$$

$$\tau^{(k)}=\frac{\sqrt{X}_i}{2\sigma^{2(k)}}\exp(Z_i\theta^{(k)}){S_0}^{(k)}.$$

Note that ${}_0F_1(1,\tau^2)=J_0(2\tau\sqrt{-1})=I_0(2\tau)$, where $J_0(z)$ is the zero-order Bessel function of first kind, $I_0(z)$ is the zero-order modified Bessel function of first kind, which satisfies

$$J_\nu'(x)=J_{\nu-1}(x)-\frac{\nu}{x}J_\nu(x),$$

and

$$J_{-n}(x)=(-1)^nJ_n(x),\quad I_n(z)=i^{-n}J_n(zi).$$

In the M-step, we maximize the parameters of the augmented log-likelihood Q from Eq. (4) w.r.t. (θ,σ^2,S_0^2). Omitting the items not depending on these parameters, Q can be expressed as

$$\begin{aligned}&\sum_{i=1}^{m}\left(\log(S_0^2)-2\log(\sigma^2)+2Z_i\theta\right)\langle N_i\rangle^{(k)}-m\log(\sigma^2)\\&-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(S_0^2\exp(2Z_i\theta)+X_i\right)\quad.\end{aligned}\tag{A.1}$$

It is easy to see in Eq. (A.1) that the log likelihood w.r.t. σ^2 and S_0^2 are inverse Gamma and Gamma distributions, respectively. Hence, we update these two parameters by their modes:

$$\widehat{\sigma^2}_{ML}:=\underset{\sigma_g^2}{\operatorname{argmax}}(Q)=\frac{\sum_{i=1}^{m}(X_i+\exp(2\hat\theta Z_i)\hat S_0{}^2)}{2\sum_{i=1}^{m}(2\langle N_i\rangle+1)}\tag{A.2}$$

and

$$\widehat{S_0^2}_{ML}:=\underset{S_0^2}{\operatorname{argmax}}(Q)=\frac{2\widehat{\sigma^2}_{ML}\sum_{i=1}^{m}\langle N_i\rangle}{\sum_{i=1}^{m}\exp(2Z_i\hat\theta)}.\tag{A.3}$$

To apply the Fisher scoring method, the score of θ is

$$\mathcal{S}(\theta)=2\sum_{i=1}^{m}\langle N_i\rangle Z_i-\frac{\widehat{S_0^2}_{ML}}{\widehat{\sigma^2}_{ML}}\sum_{i=1}^{m}\exp(2Z_i\theta)Z_i,\tag{A.4}$$

and the Fisher-information is given by

$$J(\theta)=E\left[-\frac{\partial^2Q}{\partial\theta_h\partial\theta_k}\right]=\frac{\widehat{S_0^2}_{ML}}{\widehat{\sigma^2}_{ML}}\sum_{i=1}^{m}\exp(2Z_i\theta)Z_iZ_i^T.\tag{A.5}$$

Appendix B. Maximization of Rician log-likelihood

Without data agumentation, we have to directly maximize the Rician log-likelihood Q_{Rician}, in short Q_r thereafter, by using some typical MLE method, such as gradient descent. Then the first (the score) and second derivatives of Q_r are usually required. The log-likelihood Q_r is

$$\begin{aligned}Q_r\ =\ &\text{const.}-m\log(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(Y_i^2+\exp(2Z_i\theta)S_0^2\right)\\&+\sum_{i=1}^{m}\log I_0\left(\frac{Y_i\exp(Z_i\theta)\sqrt{S_0^2}}{\sigma^2}\right),\end{aligned}$$

where $I_k(\tau)$ are modified Bessel functions of first kind satisfying

$$I_0'(\tau)=I_1(\tau),\quad I_0''(\tau)=I_1'(\tau)=(I_0(\tau)+I_2(\tau))/2.$$

The score of σ^2 and S_0^2 are respectively given by

$$\begin{aligned}\frac{\partial Q_r}{\partial\sigma^2}\ =\ &-\frac{m}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{m}\left(Y_i^2+\exp(2Z_i\theta_i)S_0^2\right)\\&-\frac{1}{\sigma^4}\sum_{i=1}^{m}g\left(Y_i\exp(Z_i\theta)S_0\sigma^{-2}\right)Y_i\exp(Z_i\theta)S_0\end{aligned}\tag{B.1}$$

and

$$\begin{aligned}\frac{\partial Q_r}{\partial S_0^2}\ =\ &-\frac{1}{\sigma^2}\sum_{i=1}^{m}\exp(2Z_i\theta_i)\\&+\frac{1}{2\sigma^2\sqrt{S_0^2}}\sum_{i=1}^{m}g\left(Y_i\exp(Z_i\theta)S_0\sigma^{-2}\right)Y_i\exp(Z_i\theta).\end{aligned}\tag{B.2}$$

The score of θ is given by

$$\begin{aligned}\frac{\partial Q_r}{\partial\theta_k}\ =\ &-\frac{S_0^2}{\sigma^2}\sum_{i=1}^{m}\exp(2Z_i\theta_i)Z_{ik}\\&+\frac{1}{\sigma^2}\sum_{i=1}^{m}g\left(Y_i\exp(Z_i\theta)S_0\sigma^{-2}\right)Y_i\exp(Z_i\theta)S_0Z_{ik}.\end{aligned}\tag{B.3}$$

The Hessian of θ is

$$\begin{aligned}\frac{\partial Q_r^2}{\partial\theta_h\partial\theta_k} &= -\frac{2S_0^2}{\sigma^2}\sum_{i=1}^{m}\exp(2Z_i\theta_i)Z_{ih}Z_{ik}\\ &\quad+\frac{S_0}{\sigma^2}\sum_{i=1}^{m}Y_i\exp(Z_i\theta)Z_{ik}Z_{ih}\left\{g\left(Y_i\exp(Z_i\theta)S_0\sigma^{-2}\right)\right.\\ &\quad\left.+g'\left(Y_i\exp(Z_i\theta)S_0\sigma^{-2}\right)\frac{Y_i\exp(Z_i\theta)S_0}{\sigma^2}\right\}\\ &=\sum_{i=1}^{m}Z_{ih}Z_{ik}\left(-4t_i^2+\tau_i(g(\tau_i)+\tau_i g'(\tau_i)\right)\\ &=\sum_{i=1}^{m}Z_{ih}Z_{ik}\left(-4t_i^2+\tau_i^2-\tau_i^2\left(\frac{I_1(\tau_i)}{I_0(\tau_i)}\right)^2\right).\end{aligned}$$

where we use

$$t_i=\frac{S_0^2\exp(2Z_i\theta_i)}{2\sigma^2},\quad \tau_i=\frac{Y_i\exp(Z_i\theta)S_0}{2\sigma^2},\quad g(\tau)=\frac{d}{d\tau}\log I_0(\tau)=\frac{I_1(\tau)}{I_0(\tau)},$$

$$\begin{aligned}g'(\tau)&=\frac{d^2}{d\tau^2}\log I_0(\tau)=\frac{1}{2}\left(1+\frac{I_2(\tau)}{I_0(\tau)}\right)-\left(\frac{I_1(\tau)}{I_0(\tau)}\right)^2\\ &=1-\frac{I_1(\tau)}{\tau I_0(\tau)}-\left(\frac{I_1(\tau)}{I_0(\tau)}\right)^2\end{aligned}$$

with

$$I_2(\tau)=I_0(\tau)-\frac{2I_1(\tau)}{\tau}.$$

For SNR > 10, the corresponding Fisher-information matrix is approximated by

$$\mathcal{I}_r(\theta)=E\left[-\frac{\partial Q_r^2}{\partial\theta_h\partial\theta_k}\right]\approx\sum_{i=1}^{m}Z_{ih}Z_{ik}\left(\frac{S_0^2}{\sigma^2}\exp(2Z_i\theta)-\frac{1}{2}\right),\qquad(B.4)$$

where (see Andersson, 2008)

$$E\left[\tau_i^2\left(\frac{I_1(\tau_i)}{I_0(\tau_i)}\right)^2\right]\approx\left(\frac{S_0^2}{\sigma^2}\exp(2Z_i\theta)\right)^2+\frac{S_0^2}{\sigma^2}\exp(2Z_i\theta)-\frac{1}{2}.$$

Appendix C. Theory of the EM algorithm by the phase data

Consider the Rician noise model in Eq. (1), and define the phase

$$\varphi:=arg\,(S+\varepsilon_1+i\varepsilon_2)\in[0,2\pi)$$

such that

$$S+\varepsilon_1=Y\cos(\varphi),\quad \varepsilon_2=Y\sin(\varphi).$$

It follows from the Bayes formula that the joint density of φ and Y for fixed S and σ^2 is given by

$$\begin{aligned}&p_{S,\sigma^2}(y,\varphi)\frac{y}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}(Y\cos(\varphi)-S)^2-\frac{1}{2\sigma^2}Y^2\sin(\varphi)^2\right)\\ &=\frac{y}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}(Y^2+S^2-2SY\cos(\varphi))\right)=p_{S,\sigma^2}(y)p_{S,\sigma^2}(\varphi|Y),\end{aligned}$$

or alternatively, similar formula can be found in Koay and Basser (2006) without using the Bayes theorem. Here the conditional density

$$p_{S,\sigma^2}(\varphi|Y)=\frac{1}{2\pi I_0(SY/\sigma^2)}\exp\left(\frac{SY}{\sigma^2}\cos(\varphi)\right),\quad\varphi\in[0,2\pi),\qquad(C.1)$$

is an instance of the symmetric Von Mises distribution on the circle. See Section 4.3.2. in Fisher et al. (1987). Note also that if the data $Y=0$, we get we get a Gaussian likelihood

$$p_{S,\sigma^2}(\varepsilon_r=-S,\varepsilon_i=0)=\frac{y}{2\pi\sigma^2}\exp\left(-\frac{S^2}{2\sigma^2}\right),$$

and in such a case the augmentation is not needed.

Appendix D. EM with latent phase measurements in multicompartment models

Zhu et al. (2013) introduces a related EM algorithm based on data augmentation with the complete complex-valued measurements $\mathbf{Y}=(Y_{ij}:1\le i\le N,\,1\le j\le M)$ for the individual diffusion compartments, and incomplete magnitude measurements $S_i=|\sum_j Y_{ij}|$. The E-step gives

$$\begin{aligned}Q(\Theta|\Theta^{(k)})&=E\left[\ell(\Theta|Y)|S,\Theta^{(k)}\right]=\int\ell(\Theta|\mathbf{Y})p(\mathbf{Y}|\mathbf{S},\Theta^{(k)})d\mathbf{Y}\\ &=\text{const.}-(M+1)N\quad\log(\sigma^2)\\ &\quad+\frac{M+1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=0}^{M}E\left[2\nu_{ij}\Re(Y_{ij})-|Y_{ij}|^2-\nu_{ij}^2|S_i,\Theta^{(k)}\right]\end{aligned}$$

where $\nu=\nu(\Theta)$ and $\Re(z)$ denotes the real part of a complex z. Since

$$\begin{aligned}E\left[\Re(Y_{ij})|S_i,\Theta^{(k)}\right]&=E\left[E\left[\Re(Y_{ij})|\Re(Y_i)\right]|S_i,\Theta^{(k)}\right]\\ &=\nu_{ij}^{(k)}+\frac{1}{M+1}E\left[\Re(Y_i)|S_i,\Theta^{(k)}\right]-\frac{\nu_i^{(k)}}{M+1},\\ \text{and}\quad E\left[\Re(Y_i)|S_i,\Theta^{(k)}\right]&=\frac{S_iI_1\left(S_i\nu_i^{(k)}/\sigma^2\right)}{I_0\left(S_i\nu_i^{(k)}/\sigma^2\right)},\end{aligned}$$

where $\nu^{(k)}=\nu(\Theta^{(k)})$, we obtain up to additive and multiplicative constants which do not depend on Θ, we obtain Eq. (7) in Zhu et al. (2013):

$$\begin{aligned}&Q(\Theta|\Theta^{(k)})=\text{const.}\\ &+\text{const.}\sum_{i,j}\left\{2\nu_{ij}\left(\frac{S_i}{M+1}\frac{I_1\left(S_i\nu_i^{(k)}/\sigma^2\right)}{I_0\left(S_i\nu_i^{(k)}/\sigma^2\right)}-\frac{\nu_i^{(k)}}{M+1}+\nu_{ij}^{(k)}\right)-\nu_{ij}^2\right\}.\end{aligned}\qquad(D.1)$$

In the M-step it is used the gradient of (D.1), given by

$$\begin{aligned}&\frac{\partial Q(\Theta|\Theta^{(k)})}{\partial\theta}\\ &=\text{const.}\sum_{i,j}\left(\frac{S_i}{M+1}\frac{I_1\left(S_i\nu_i^{(k)}/\sigma^2\right)}{I_0\left(S_i\nu_i^{(k)}/\sigma^2\right)}-\frac{\nu_i^{(k)}}{M+1}+\nu_{ij}^{(k)}-\nu_{ij}\right)\frac{\partial\nu_{ij}(\Theta)}{\partial\theta}\end{aligned}$$

We note that one could use simply the EM algorithm with for a single component ($M=0$) with latent data ($Y_i:i=1,\ldots,n$), optimizing in the M-step

$$Q(\Theta|\Theta^{(k)})=\text{const.}+\text{const.}\sum_i\left\{\frac{2\nu_iS_iI_1\left(S_i\nu_i^{(k)}/\sigma^2\right)}{I_0\left(S_i\nu_i^{(k)}/\sigma^2\right)}-\nu_i^2\right\}$$

with gradient

$$\frac{\partial Q(\Theta|\Theta^{(k)})}{\partial\theta} = \text{const.}\sum_i \left\{ \frac{S_i I_1\left(S_i\nu_i^{(k)}/\sigma^2\right)}{I_0\left(S_i\nu_i^{(k)}/\sigma^2\right)} - \nu_i \right\} \frac{\partial\nu_i(\Theta)}{\partial\theta}$$

Since the phase augmentation under the single compartment model is quite similar in the computation by applying the proposed EM-MLE scheme, therefore, it is straightforward to extend our methods to the multiple compartment case.

References

Andersson JL. Maximum a posteriori estimation of diffusion tensor parameters using a Rician noise model: why, how and but. Neuroimage 2008];42(4):1340–56.

Barber P, Darby D, Desmond P, Yang Q, Gerraty R, Jolley D, et al. Prediction of stroke outcome with echoplanar perfusion- and diffusion-weighted MRI. Neurology 1998];51(2):418–26.

Barmpoutis A, Vemuri BC, Shepherd TM, Forder JR. Tensor splines for interpolation and approximation of DT-MRI with applications to segmentation of isolated rat hippocampi. IEEE Trans Med Imaging 2007];26(11):1537–46.

Barmpoutis A, Hwang MS, Howland D, Forder JR, Vemuri BC. Regularized positive-definite fourth order tensor field estimation from DW-MRI. NeuroImage 2009];45(1):S153–62.

Fisher NI, Lewis T, Embleton BJ. Statistical analysis of spherical data. Cambridge University Press; 1987].

Gasbarra D, Liu J, Railavo J. Data augmentation in Rician noise model and Bayesian diffusion tensor imaging; 2014] arXiv:0935897.

Ghosh A, Milne T, Deriche R. Constrained diffusion kurtosis imaging using ternary quartics and MLE. Magn Reson Med 2014];71(4):1581–91.

Griffanti L, Baglio F, Preti GM, Cecconi P, Rovaris M, Baselli G, et al. Signal-to-noise ratio of diffusion weighted magnetic resonance imaging: estimation methods and in vivo application to spinal cord. Biomed Signal Process Control 2012];7(3):285–94.

Henkelman RM. Measurement of signal intensities in the presence of noise in MR images. Med Phys 1985];12(2):232–3.

Jaynes ET. Probability theory: the logic of science. Cambridge University Press; 2003].

Jeffrey A, Zwillinger D. Table of integrals, series, and products. Academic Press; 2007].

Jones DK, Basser PJ. Squashing peanuts and smashing pumpkins: how noise distorts diffusion-weighted MR data. Magn Reson Med 2004];52(5):979–93.

Koay CG, Basser PJ. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. J Magn Reson 2006];179(2):317–22.

Koay CG, Özarslan E, Basser PJ. A signal transformational framework for breaking the noise floor and its applications in MRI. J Magn Reson 2009];197(2):108–19.

Landman B, Bazin P-L, Prince J. Diffusion tensor estimation by maximizing Rician likelihood. IEEE 2007]:1–8.

Lange K. Optimization, 2nd ed., vol. 95, Springer texts in statistics; 2013].

Leemans A, Jeurissen B, Sijbers J, Jones D. Exploredti: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. In: 17th annual meeting of international society for magnetic resonance in medicine; 2009]. p. 3537.

McCullagh P, Nelder JA. Generalized linear models; 1989].

Qi L, Yu G, Wu EX. Higher order positive semidefinite diffusion tensor imaging. SIAM J Imaging Sci 2010];3(3):416–33.

Rajan J, Jeurissen B, Verhoye M, Van Audekerke J, Sijbers J. Maximum likelihood estimation-based denoising of magnetic resonance images using restricted local neighborhoods. Phys Med Biol 2011];56(16):5221.

Solo V, Noh JJ. An EM algorithm for Rician FMRI activation detection. In: From nano to macro, ISBI 2007. 4th IEEE international symposium on biomedical imaging. IEEE; 2007]. p. 464–7.

Sparacino G, Tombolato C, Cobelli C. Maximum-likelihood versus maximum a posteriori parameter estimation of physiological system models: the c-peptide impulse response case study. IEEE Trans Biomed Eng 2000];47(6):801–11.

Veraart J, Van Hecke W, Sijbers J. Constrained maximum likelihood estimation of the diffusion kurtosis tensor using a Rician noise model. Magn Reson Med 2011];66(3):678–86.

Zhu H, Zhang H, Ibrahim JG, Peterson BS. Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data. J Am Stat Assoc 2007];102(480):1085–102.

Zhu X, Gur Y, Wang W, Fletcher P. Model selection and estimation of multicompartment models in diffusion MRI with a Rician noise model. In: Information processing in medical imaging, vol. 7917 of Lecture notes in computer science. Berlin, Heidelberg: Springer; 2013]. p. 644–55.

III

AN IMPROVED EM ALGORITHM FOR SOLVING MLE IN CONSTRAINED DIFFUSION KURTOSIS IMAGING OF HUMAN BRAIN

by

Liu, J 2019

Submitted manuscript

An improved EM algorithm for solving MLE in constrained diffusion kurtosis imaging of human brain

Jia Liu*

Department of Mathematics and Statistics, University of Helsinki P.O. Box 68 FI-00014 Finland
Department of Mathematics and Statistics, University of Jyväskylä, P.O.Box 35 (MaD) FI40014 Finland

Abstract

Background: Diffusion kurtosis imaging (DKI) as an advanced medical imaging technique extends the parametric model for diffusion tensor imaging (DTI) by including the diffusional kurtosis term which describes non-Gaussian properties of water diffusion due to micro-structural tissue barriers. The model allows the tensor parameters to be estimated constrained on the physical relevance of water diffusion, which leads to a nonlinear regression problem including nonlinear constraints in the estimation.

New methods: We propose an efficient computational method, the expectation-maximization (EM) algorithm based on the maximum likelihood estimation with constraints (CMLE) for the DKI estimation. We consider the Rician noise-corrupted signal model by introducing Von-Mises data augmentation and accommodated all the constraints in DKI. A constrained Fisher-scoring numerical method is suggested for tensor optimization. Two extended algorithms, constrained weighted the least square (CWLS) with interior method (CWLS-IP) and constrained nonlinear least squares algorithm (CWLS-LLS) are also proposed.

Results: The method improves the efficiency of the traditional Rician MLE based methods. The results show promising performance by means of conducting the proposed method both on synthetic and real data from human brain.

Comparison with Existing Methods:: We compare our EM method (EM-IP) with CWLS-IP, CWLS-LLS and the competing alternatives including the weighted least

*Corresponding author
Email address: jia.liu@helsinki.fi; jia.2.liu@jyu.fi (Jia Liu)

squares (WLS), the constrained maximum likelihood estimation with sequential quadratic programming (MLE-SQP) and constrained weighted the least square with SQP (CWLS-SQP).

Conclusions: Our EM method perform much better than the alternatives especially for data retrieved from a low regime of signal to noise ratio (SNR) and from the high b values.

Highlights.

- We originate Von Mises data augmentation in diffusion MRI and propose an efficient EM method for the DKI estimation under the Rician noise-corrupted signal model and extend other methods.
- The constraints in DKI are accommodated by means of a constrained Fisher-scoring algorithm with precise formulas of Hessians.
- We extended the heuristic algorithm and proposed a nonlinear least squares solution (CWLS-LLS) for good initial values in the optimization, and proposed CWLS-IP for fast computation.
- Our EM method shows promising performance in a wider region of SNR and for the data with high b values.

Keywords: Barrier Method, Constrained Fisher Scoring, Data Augmentation, Diffusion Kurtosis Imaging (DKI), Maximum Likelihood Estimation (MLE), Non-Gaussianity, Positivity, Rician, Ternary Quartics (TQ), Von Mises.

---

1. Introduction

Diffusion tensor imaging (DTI) introduced by [1, 2, 3] is a sophisticated diffusion magnetic resonance imaging (MRI) reconstruction technique which enables the observer to explore in vivo the structural information and geometric organization of the brain anatomy at the microscopic level. It models the three-dimensional (3D) diffusion

process by means of the low angular resolution (rank-2) tensor matrix, where the probability distribution of the water diffusion is assumed to be Gaussian. The assumption, however, is argued to diverge significantly from reality when the model is applied to genuine biological tissues. An example is the human brain containing an appendage of complex tissues rich in microstructures such as cell membranes, boundaries and other complex compartments. Evidence of the spatially diffusional non-Gaussianity is discovered in the white matter of the human brain. Tuch et al. [4, 5] propose the high angular resolution diffusion imaging (HARDI) which does not rely on the assumption of the Gaussian distribution in the diffusion. This technique has been further extended by [6] and [7] for describing the diffusivity profile by means of high rank Cartesian tensors and tensor-based spherical harmonic representation, respectively. The diffusion kurtosis imaging (DKI) [8, 9, 10] in connection with the multi-diffusional tensor imaging techniques DTI and HARDI, has recently become popular in quantifying the degree of diffusional deviation from Gaussianity. It is referred as a natural extension of DTI by adding a high angular resolution diffusional term

$$S(b) = S_0 \exp(-bD_{app} + \frac{1}{6}b^2 D_{app}^2 K_{app}), \tag{1}$$

see for instance [11, 12, 13], where S_0 is the signal intensity without diffusion weighting known as unattenuated signal, $S(b)$ is the true signal magnitude and b is the diffusion weighting amplitude or the so-called the b value,

$$D_{app} := \mathbf{g}^T D \mathbf{g} = \sum_{\ell_1,\ell_2=1}^{3} g_{\ell_1} g_{\ell_2} D_{\ell_1,\ell_2}$$

is called the apparent diffusional coefficient with pulse gradient $\mathbf{g}$ and tensor matrix D, and

$$K_{app} = \left(\frac{\overline{tr(D)}}{D_{app}}\right)^2 \sum_{\ell_1,\ell_2,\ell_3,\ell_4=1}^{3} g_{\ell_1} g_{\ell_2} g_{\ell_3} g_{\ell_4} W_{\ell_1,\ell_2,\ell_3,\ell_4}$$

is the apparent diffusion kurtosis. Here "tr" denotes the trace of the matrix operator, and $\overline{tr(D)} = 1/3 \sum_{i=1}^{3} tr(D_{ii})$. The definition of the kurtosis tensor $W_{\ell_1,\ell_2,\ell_3,\ell_4}$ can be found for instance in [8]. The model contains implicitly three constraints (see also [14, 15]) which are:

1. The physical relevance and biological plausibility require that D is positive definite.

2. $K_{app} \geq 0$ is most likely the lower bound constraint of the apparent diffusion kurtosis, although in theory $K_{app} \geq -2$. It further implies that the fourth symmetric kurtosis tensor W should be positive definite in three dimensions (3D). Further, $K_{app} \geq 0$ requires $D_{app} \geq 0$, which in general is guaranteed by # 1.

3. The upper bound constraint is $K_{app} \leq 3/(bD_{app})$.

The first two constraints originate from the physical relevance that the diffusivity function should be positive. The third constraint is inherited from the assumption that the signal intensity $S(b)$ is a monotonically decreasing function in the b amplitudes. The model thus can only utilize the b value less than 3000 s/mm^2. When comparing DTI and HARDI, DKI additionally requires at least three distinct b values and 15 diffusional directions, but may bring auxiliary information of diffusional heterogeneity which could contribute to the diagnosis of neuropathologies [16, 17, 10]. Although DKI has limitations in connection with the b value, data captured in that region ($b \leq 3000 s/mm^2$) are much more feasible in clinical imaging protocols.

In [18], we proposed a Poisson data augmentation that works in connection of the Rician likelihood in count data space to detect the water diffusion by means of diffusion tensors using both DTI and HARDI. The methodology is designed especially for data retrieved from the low regime of signal to noise ratio (SNR) and with high b values, though the algorithm did not consider the positivity of the tensor matrices with DTI. This work continues [18], introducing an alternative idea of data augmentation, that is, the *Von Mises* data augmentation to solve the regression problem with Rician likelihood in diffusion MRI. In doing that we can directly work with generalized linear modeling (GLM) under the Rician noise model in the phase data space. We propose a new expectation-maximization (EM) algorithm that considers all constraints with an advanced model of diffusion, diffusion kurtosis imaging. A constrained stabilized Fisher-scoring algorithm is proposed in connection with the Von Mises data augmentation to obtain optimal solution and fast convergence in the estimation of the tensor parameters. The method, therefore, can also be applied for DTI and HARDI

straightforwardly. Additionally, we use the barrier method, see for instance [19], to deal with the nonlinear constraint # 3. To sum up, in this work we propose an efficient computational method with DKI to solve a nonlinear constrained regression problem. The comparison of the methods, such as precision as a function of the b value measured in terms of mean square error, precision as a function of signal to noise ratio measured in terms of mean square error and computational time per voxel have been well demonstrated in the simulation study.

The paper is structured as follows: Section 2 reviews the Rician noise model, generalized DTI and some general ideas of the positivity constraints. In Section 3, we give an exposition of Von Mises data augmentation and illustrate how it works with DKI. In the following section, we focus on the EM and the constrained Fisher-scoring algorithms in DKI. The proposed methods have been implemented on both synthetic and real data in Section 5 and 6, respectively with a discussion in Section 7.

2. Rician likelihood and constrained DKI

2.1. MR noise and Rician magnitude

We consider first the model for an observation in a single voxel with a given acquisition. The noise ε in the raw MR-acquisition model is composed of two $i.i.d.$ Gaussian random variables, ε_r and ε_i, with zero mean μ and common variance σ^2, which originate from the real and imaginary components, respectively. The joint density of the MR noise is expressed as $p_{\mu,\sigma^2}(\varepsilon_r,\ \varepsilon_i) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\varepsilon_r^2+\varepsilon_i^2}{2\sigma^2}\right)$, see also [20]. The observation is the corrupted signal intensity (corrupted by the complex-valued noise) and is defined as $Y = \sqrt{(S+\varepsilon_r)^2+\varepsilon_i^2}$. It has the Rician distribution with the probability density function

$$p_{S,\sigma^2}(y) = \frac{y}{\sigma^2}\exp\left(-\frac{y^2+S^2}{2\sigma^2}\right) I_0\left(\frac{yS}{\sigma^2}\right)\mathbb{1}(y \geq 0), \tag{2}$$

[21, 22, 23], where $S \in \mathbb{R}^+$ is the magnitude of the true (noise-free) signal, $I_\alpha(\cdot)$ is the α-order modified Bessel function of the first kind, and $\mathbb{1}(\cdot)$ is the indicator function.

2.2. Generalized DTI and Constrained DKI

DTI is a simple but an elegant technique to model the diffusion. The model is simply expressed as $S(b) = S_0 \exp(-b\mathbf{g}^T D\mathbf{g})$, where b states the multi-shells of b value and $\mathbf{g}$ indicates the gradient vector from the unit sphere. For rank n ($n > 2$ and $n \in 2\mathbb{N}$) tensor, the model can be extended as

$$S(b) = S_0 \exp(-bD^{(n)}_{app}), \text{ and } D^{(n)}_{app} := \sum_{\ell_1=1}^{3}\sum_{\ell_2=1}^{3}\cdots\sum_{\ell_n=1}^{3} D_{\ell_1,\ell_2,\ldots,\ell_n} g_{\ell_1} g_{\ell_2} \cdots g_{\ell_n},$$

which is referred as Generalized DTI [7]. It formulates the diffusivity profile by means of a high rank $n > 2$ Cartesian tensor in HARDI. We can further parametrize $-bD^{(n)}_{app} = Z\theta$ that results in a nonlinear regression model of diffusive signal attenuation $S = S_0 \exp(Z(b,\mathbf{g})\theta)$, where θ is the tensor parameter and Z denotes a design matrix.

The observations in a single voxel are obtained under of a chosen design matrix that consists of m acquisitions. In DTI, the six distinct elements of D are defined as $\theta_D = (\theta_1,\ldots,\theta_6)^\top := \big(D_{11}, D_{22}, D_{33}, D_{12}, D_{13}, D_{23}\big)^\top$. The corresponding design matrix, composed of m acquisitions is given by

$$Z_D = Z(b,\mathbf{g}) = -b \begin{pmatrix} g^2_{11} & g^2_{21} & g^2_{31} & 2g_{11}g_{21} & 2g_{11}g_{31} & 2g_{21}g_{31} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ g^2_{1j} & g^2_{2j} & g^2_{3j} & 2g_{1j}g_{2j} & 2g_{1j}g_{3j} & 2g_{2j}g_{3j} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ g^2_{1m} & g^2_{2m} & g^2_{3m} & 2g_{1m}g_{2m} & 2g_{1m}g_{3m} & 2g_{2m}g_{3m} \end{pmatrix}. \quad (3)$$

Using this parametrization, DKI in Eq. (1) can be further expressed as

$$\begin{aligned} &S(b) = \\ &S_0 \sum_{j=1}^{m} \exp\Bigg(-b \sum_{\ell_1,\ell_2=1}^{3} g_{\ell_1} g_{\ell_2} D_{\ell_1,\ell_2} + \frac{b^2}{6} \Big(\sum_{\ell_1=1}^{3} \frac{D_{\ell_1\ell_1}}{3}\Big)^2 \\ &\sum_{\ell_1,\ell_2,\ell_3,\ell_4=1}^{3} g_{\ell_1} g_{\ell_2} g_{\ell_3} g_{\ell_4} W_{\ell_1,\ell_2,\ell_3,\ell_4} \Bigg) = S_0 \sum_{j=1}^{m} \exp(Z_{D_j}\theta_D + Z_{W_j}\theta_W(\overline{tr(D)}^2; W)), \end{aligned} \quad (4)$$

where the jth row of the design matrix $Z_W \in \mathbb{R}^{m\times 15}$ is $Z_{W_j} = \frac{b^2}{6}(g^4_{1j}, g^4_{2j}, g^4_{3j}, 6g^2_{1j}g^2_{2j}, 6g^2_{1j}g^2_{3j}, 6g^2_{2j}g^2_{3j}, 12g^2_{1j}g_{2j}g_{3j}, 12g_{1j}g^2_{2j}g_{3j}, 12g_{1j}g_{2j}g^2_{3j}, 4g^3_{1j}g_{2j}, 4g^3_{1j}g_{3j}, 4g^3_{2j}g_{1j}, 4g^3_{2j}g_{3j}, 4g^3_{3j}g_{1j}, 4g^3_{3j}g_{2j}), j = 1\cdots m.$

2.3. Constrained DKI

Since D is a 3×3 symmetric positive definite matrix, by the Cholesky decomposition we have $D = UU^T$ and $U = \begin{pmatrix} L_1 & & \\ L_4 & L_2 & \\ L_5 & L_6 & L_3 \end{pmatrix}$, which is a lower triangular matrix. Without changing the design matrix Z_D, the tensor parameter θ_D can be reparametrized as a function of $L = (L_1, L_2, L_3, L_4, L_5, L_6)$ so that
$\theta_D(L) = (L_1^2, L_2^2 + L_4^2, L_3^2 + L_5^2 + L_6^2, L_1L_4, L_1L_5, L_4L_5 + L_2L_6)$.
The corresponding Jacobian matrix is

$$J_L = \frac{\partial \theta_D}{\partial L_{j=1,\cdots,6}} = \begin{pmatrix} 2L_1 & & & & & \\ & 2L_2 & & 2L_4 & & \\ & & 2L_3 & & 2L_5 & 2L_6 \\ L_4 & & & L_1 & & \\ L_5 & & & & L_1 & \\ & L_6 & & L_5 & L_4 & L_2 \end{pmatrix}. \tag{5}$$

The constraint # 2 (see page 3) implies that W_{app} is non-negative. We apply the ternary quartic (TQ) to guarantee the positivity condition as in [14], and express the non-negative apparent kurtosis coefficients as

$$W_{app} = \sum_{i=1}^{3} \left(\mathbf{v}^T \mathbf{q}_i \right)^2 = \mathbf{v}^T Q Q^T \mathbf{v} = \mathbf{v}^T G \mathbf{v}, \tag{6}$$

where $\mathbf{v} = [g_1^2, g_2^2, g_3^2, g_1g_2, g_1g_3, g_2g_3]^T$, and $Q = [\mathbf{q}_1|\mathbf{q}_2|\mathbf{q}_3]$ is a 6×3 matrix, containing three 6×1 vectors $\mathbf{q}_i$. The Gram matrix $G = QQ^T$ is a 6×6 positive symmetric matrix composed of the all fifteen kurtosis tensor elements plus six free parameters (see [24] for details). Let $\theta_Q := \overline{tr(D)} \begin{pmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \mathbf{q}_3 \end{pmatrix}$, and $P_j = \frac{b^2}{6} \begin{pmatrix} \mathbf{vv^T} & & \\ & \mathbf{vv^T} & \\ & & \mathbf{vv^T} \end{pmatrix}$ be an 18×18 matrix at the signal acquisition j. Then Eq. (4) can be written as

$$S = S_0 \sum_{j=1}^{m} \exp\left(Z_{D_j} \theta_D(L) + \theta_Q^T P_j \theta_Q \right). \tag{7}$$

3. Data augmentation

The idea of data augmentation (DA) arises from the missing data model where the likelihood is of the form

$$g(y|\theta) = \int_n f(y,n|\theta)dn,$$

see for instance [25, 26, 27]. The model contains missing data N. The likelihood $f(y,n|\theta)$ is called the *complete-data* likelihood. DA relies on the idea that in some instances it is easier to deal with $f(y,n|\theta)$ than with $g(y|\theta)$. In this work, we propose a DA scheme, an alternative to [18], in order to ease the computational problem in connection with the nonlinear regression model of Eq. (2).

3.1. Von Mises data augmentation

The Rician likelihood in Eq. (2) for a given signal can be represented in the phase data space through a transformation from the real and imaginary images to the arctangent of their ratio, see [20, 23, 28].

Let φ be the phase data

$$\varphi := \arg\Big(S + \varepsilon_r + i\varepsilon_i\Big) \in [0, 2\pi)$$

such that

$$S + \varepsilon_r = Y\cos(\varphi), \quad \varepsilon_i = Y\sin(\varphi).$$

By change of variables, the joint density of φ and Y for fixed S and σ^2 is given by

$$\begin{aligned} p_{S,\sigma^2}(y,\varphi) &= \frac{y}{2\pi\sigma^2}\exp\bigg(-\frac{1}{2\sigma^2}\big(y\cos(\varphi)-S\big)^2 - \frac{1}{2\sigma^2}y^2\sin(\varphi)^2\bigg) \\ &= \frac{y}{2\pi\sigma^2}\exp\bigg(-\frac{1}{2\sigma^2}\big(y^2+S^2-2yS\cos(\varphi)\big)\bigg) \\ &= p_{S,\sigma^2}(y)p_{S,\sigma^2}(\varphi|y). \end{aligned} \tag{8}$$

Then we have:

1. Eq. (8) is the Rician noise model presented in the phase data space.

2. The conditional density

$$p_{S,\sigma^2}(\varphi|y) = \frac{1}{2\pi I_0(yS/\sigma^2)} \exp\left(\frac{yS}{\sigma^2}\cos(\varphi)\right), \quad \varphi \in [0, 2\pi), \tag{9}$$

is an instance of the symmetric Von Mises distribution on the circle, see Chapter 4.3.2 in [29]. Note also that for $y = 0$ we obtain the Gaussian likelihood

$$p_{S,\sigma^2}(\varepsilon_r = -S, \varepsilon_i = 0) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\varepsilon_r^2}{2\sigma^2}\right),$$

in which case augmentation is not needed.

3. Using the parametrization $S = \exp(Z\theta)$, the complete-data likelihood

$$f(y, \varphi|\theta, Z) = \frac{y}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\left(y^2 + \exp(2Z\theta) - 2\exp(Z\theta)y\cos(\varphi)\right)\right) \tag{10}$$

w.r.t. θ has a Gaussian response up to a constant depending on the observation and a log link function

$$g(\mathbb{E}(Y, \varphi|\theta, Z)) = \log(\mu) = Z\theta$$

conditionally on $\cos\varphi$, and the mean of the response, $\mu = \exp(Z\theta)$.

Consequently, the nonlinear regression problem in Eq. (2) has been transferred into the generalized linear modeling (GLM) [30] framework augmenting the data y by the phase data φ.

4. EM in the constrained DKI

In the maximum likelihood estimation (MLE), we employ the EM algorithm in connection with data augmentation to find the optimal solutions of the unknown parameters in the constrained DKI. The theory of the EM algorithm can be found, for example, in [31, 32, 33]. After initializing the parameters, the EM algorithm typically proceeds in two steps: in the E-step we calculate, under the current parameter values, the expectation of the log-likelihood w.r.t. the conditional distribution of the latent variable given the observations; in the M-step, we update the parameters by maximizing the complete log-likelihood. We name the proposed method EM-IP, because we apply the interior point method in the optimization.

In detail, we directly work with the complete-data log-likelihood in Eq. (10) for m acquisitions in the design matrix by introducing a Von Mises distributed latent phase variable φ. After omitting the constant, the complete-data log-likelihood of the constrained DKI under the Rician likelihood model is given by

$$m\log(\sigma^{-2}) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}\Bigg\{Y_j^2 + S_0^2\exp\Big(2Z_{D_j}\theta_D + 2\theta_Q^T P_j\theta_Q\Big) - 2\cos(\varphi_j)Y_jS_0\exp\Big(Z_{D_j}\theta_D + \theta_Q^T P_j\theta_Q\Big)\Bigg\}, \tag{11}$$

where the sum is over the m acquisitions in a voxel.
To simplify the notations, we define $\zeta_j^{(k)} := \exp(Z_{Dj}\theta_D^{(k)})$, $\psi_j^{(k)} := \exp\Big((\theta_Q^{(k)})^T P_j\theta_Q^{(k)}\Big)$ and $\tau_j^{(k)} := Y_j\big\langle\cos(\varphi_j)\big\rangle^{(k)}$, where $<\cdot>$ is a shorthand notation for the expectation.
In the E-step, given the current parameter estimates $(\theta_D^{(k)}, \theta_Q^{(k)}, S_0^{(k)}, (\sigma^2)^{(k)})$, we update for each acquisition j the conditional expectation of $\cos\varphi_j$ given data Y_j using Eq. (9), and obtain

$$\big\langle\cos\varphi_j\big\rangle^{(k)} = \frac{I_1\Big(Y_jS_0^{(k)}\zeta_j^{(k)}\psi_j^{(k)}(\sigma^{-2})^{(k)}\Big)}{I_0\Big(Y_jS_0^{(k)}\zeta_j^{(k)}\psi_j^{(k)}(\sigma^{-2})^{(k)}\Big)} \in [-1,1]. \tag{12}$$

This formula is fairly easy to derive from the first moment of the Von Mises distribution, see Appendix B.

The likelihood in Eq. (11) w.r.t. the parameters σ^{-2} (known as the profile likelihood) is the Gamma likelihood with shape and rate parameters given by $(m+1)$ and $\frac{1}{2}\Big\{Y_j^2 + (S_0^{(k)})^2(\zeta_j^{(k)})^2(\psi_j^{(k)})^2 - 2S_0^{(k)}\tau_j^{(k)}\zeta_j^{(k)}\psi_j^{(k)}\Big\}$, respectively. The profile likelihood of S_0 is Gaussian with mean $\sum_{j=1}^m \tau_j^{(k)}\zeta_j^{(k)}\psi_j^{(k)} / \sum_{j=1}^m(\zeta_j^{(k)})^2(\psi_j^{(k)})^2$.
In the M-step, we update S_0 and σ^2 by their modes in the recursions according to

$$S_0^{(k+1)} = \frac{\sum_{j=1}^m \tau_j^{(k)}\zeta_j^{(k)}\psi_j^{(k)}}{\sum_{j=1}^m(\zeta_j^{(k)})^2(\psi_j^{(k)})^2} \tag{13}$$

and

$$(\sigma^{-2})^{(k+1)} = 2m/\sum_{j=1}^{m}\Big\{Y_j^2 + (S_0^{(k)})^2(\zeta_j^{(k)})^2(\psi_j^{(k)})^2 - 2S_0^{(k)}\tau_j^{(k)}\zeta_j^{(k)}\psi_j^{(k)}\Big\}, \tag{14}$$

respectively.

4.1. Constrained Fisher scoring (CFS)

In the estimation, the marginal log-likelihoods of the parameters L or θ_Q derived in Eq. (11) are not of the standard form that result in intractable computational problems. To solve the problem we use the Laplace approximation for the profile log-likelihood $f(L|y,\varphi,Z)$ and $f(\theta_D|y,\varphi,Z)$ and propose a stabilized constrained Fisher scoring algorithm for optimizing the two functions.

Define $\Theta := (L,\theta_Q)$. After omitting the terms in Eq. (11) which do not depend on L and θ_Q, the maximization of the log-likelihood with constraints is reduced to minimizing

$$f(\Theta) := \frac{1}{2\sigma^2}\sum_{j=1}^{m}\left\{(S_0)^2(\zeta_j)^2(\psi_j)^2 - 2S_0\tau_j\zeta_j\psi_j\right\}, \tag{15}$$

with the constraint

$$g(\Theta) := 2\theta_Q^T P_j\theta_Q + Z_{D_j}\theta_D(L) \leq 0,$$

where the nonlinear constraint $g(\Theta)$ is derived from $K_{app} \leq 3/(bD_{app})$ so that zero b value can be also considered in the estimation.

4.2. Regularization

The Fisher scoring (FS) can sometimes ease dramatically the computation. For example, in our case when updating L, the computation of Fisher information is much easier than to compute directly the second derivatives, see Appendix C for details. The updating scheme of the FS may be written as

$$\Theta \leftarrow \Theta + \alpha\mathbb{S}(\Theta)\mathscr{I}(\Theta)^{(-1)}, \tag{16}$$

where $\alpha \in (0,1)$ is the step parameter. The Fisher score is $\mathbb{S}(\Theta) = \bigtriangledown f(\Theta)$. When we impose the third constraint $g(\Theta)$ on the nonlinear problem in Eq. (15), the score of Θ becomes $\mathbb{S}_\Theta := \mathbb{S}(\Theta,\lambda) \bigtriangledown f(\Theta) + \sum_{j=1}^{m} \bigtriangledown\lambda_j g(\Theta)$ and the corresponding Fisher information is

$$\mathscr{I}(\Theta) = \begin{pmatrix} \mathscr{I}(L,\lambda) & \\ & \mathscr{I}(\theta_Q,\lambda) \end{pmatrix}.$$

In order to avoid the singularity or the ill-condition of the Fisher information, we apply the Levenberg-Marquardt (LM) method, combining the gradient descent method and the Fisher-scoring method to stabilize the algorithm. Updating Θ then is

$$\Theta \to \Theta + \left[\mathscr{I}(\Theta) + \gamma \mathbb{1}\right]^{-1} \mathbb{S}(\Theta), \tag{17}$$

where $\mathbb{1}$ is an identity matrix. Further, we choose the LM parameter $\gamma = \|\mathbb{S}(\Theta)\|$, being among a few optimal choices of the Levenberg-Marquart parameter, see [34, 35], where $\|\cdot\|$ is the norm. To speed up the computation, we only apply this regularization scheme, when $\mathscr{I}(L^{(k)}, \lambda^{(k)})$ and $\mathscr{I}(\theta_Q^{(k)}, \lambda^{(k)})$ are singular or close to singular.

CFS can be achieved by means of many optimization algorithms, which are modifications of the Newton method, where the constraints are considered in the calculations. In this work we apply the barrier method to solve the constrained nonlinear optimization problem represented in Eq. (15).

4.3. Constrained weighted least squares (CWLS) by CFS, CWLS-IP

To extend the idea, we adopt CFS and the regularization scheme as presented in Section 4.1 and 4.2 in the most commonly used computational method in diffusion MRI, the weighted least squares (WLS), for estimating Θ in the constrained DKI. We apply the interior penalty algorithm for the optimization and named the method as CWLS-IP.

The Fisher information is then

$$\mathscr{I}_{cwls}(\Theta, \lambda) = \begin{pmatrix} \mathscr{I}_{cwls}(L, \lambda) & \\ & \mathscr{I}_{cwls}(\theta_Q, \lambda), \end{pmatrix}$$

where

$$\mathscr{I}_{cwls}(L,\lambda) = -J_L^T\left(\sum_{j=1}^{m} w_j\left(\log Y_j - \log S_0 - Z_{D_j}\theta_D(L) - \theta_Q P_j \theta_Q\right) Z_{Dj}^T Z_{Dj}\right) J_L$$
$$- \sum_{j=1}^{m} \lambda_j M_j,$$

and

$$\mathscr{I}_{cwls}(\theta_Q,\lambda) = -4\sum_{j=1}^{m} w_j\left(\log Y_j - \log S_0 - Z_{D_j}\theta_D(L) - \theta_Q P_j \theta_Q\right)\theta_Q^T P_j^T\,\theta_Q P_j$$
$$+2\sum_{j=1}^{m} w_j\left(\log Y_j - \log S_0 - Z_{D_j}\theta_D(L) - \theta_Q P_j \theta_Q\right) P_j - 4\sum_{j=1}^{m}\lambda_j P_j.$$

5. Constrained nonlinear least squares algorithm (CWLS-LLS)

We extends the heuristic algorithm proposed by [13], using a simply way to obtain the estimates that almost satisfy all the constraints in DKI, and meanwhile we get good and reasonable starting values in the proposed EM algorithm. In other words, the estimates $\Theta^{(0)}$ of parameter $\Theta = (L, \theta_Q)$ by CWLS-LLS fit the constraints in DKI and are as close to the WLS estimates as possible. The positive constraints by CWLS-LLS are controlled directly by the quadratic model using TQ in Eq. (6).

Let us take the estimates ($\log\widehat{S}_0$, $\widehat{\theta}_D$, $\widehat{\theta}_w$) by WLS;

1. Check the eigenvalues ($\Lambda_i, i = \{1,\cdots,3\}$) of the 3×3 tensor matrices, $D(\widehat{\theta}_D)$:

   If any $\Lambda_i \le 0$, set it to have a negligible constant value, $a \in R^+$.

   If Λ_i is changed, reconstruct $D(\widehat{\theta}_D)$ and update θ_D, fix θ_D and $\log S_0$ and then apply WLS to update θ_W and σ^2.

2. Set $\kappa = Z_W\widehat{\theta}_W$, and

   $$\text{if any } \kappa_j > -\frac{1}{2}Z_{D_j}\widehat{\theta}_D, \text{ set } \kappa_j = -\frac{1}{2}Z_{D_j}\widehat{\theta}_D.$$

   If any κ_j is changed, θ_W is known via new κ and apply WLS to update ($\log S_0, \theta_D$ and σ^2).

3. (Option) Repeat Step 1 and 2 a few iterations (e.g., 10-100) to increase the accuracy of the estimates that satisfy the constraints if any Λ_i is updated by Step 2.

4. Apply the nonlinear least squares method (LLS) on

$$\widehat{\kappa} = \frac{b^2}{6} \sum_{j=1}^{3} \left(\mathbf{v}^T \mathbf{q}_i^* \right)^2$$

to obtain $\theta_Q^{(0)}$, where $Q^* = \overline{tr(D)}Q$, see Eq. (6). Apply the Cholesky decomposition to get $L^{(0)}$.

Step 1 results in # 1 to be satisfied and preserve the directions of the positive curvature in the original tensor matrices as much as possible. Step 2 is used to reach # 3 at each acquisition. Step 3 intends to increase the accuracy of the estimates that satisfy the constraints 1 and 2. Step 4 is using TQ to meet #2, where a simple option of the initial values of LLS can be the square root of $\widehat{\theta}_W$ obtaining the first 15 elements of $Q^{(0)}$, and the remaining three elements are from random samples.

6. Simulation and case studies

We use both synthetic and real data to implement the proposed methods, the new EM algorithms (EM-IP), CWLS by IP (CWLS-IP) and CWLS-LLS, and compare with other popular alternatives: WLS, CWLS-SQP and the constrained MLE algorithm by sequential quartic programming algorithm (SQP, MLE-SQP) proposed in [14] through their results. We implemented MLE-SQP using our parametrization described in Section 2.3 directly on the Rician noise model with likelihood function of Eq. (2), where we fixed σ^2 and S_0 and applied the `fmincon` with SQP as described in [14] with the target, the minus logarithmic likelihood for optimizing Θ. All the implementation were carried out in the MATLAB environment with version R2018a. Proposal scaling is needed for dealing with numerical problems encountered with Rician noise model.

6.1. Simulation 1

In this study we simulate the (noise-free) MR signals using three models: Model 1. $S(b) = \exp(-\frac{1}{2} b D_{app})$, Model 2. $S(b) = \exp(-\frac{1}{4} b D_{app})$ and Model 3.

$S(b) = \exp(-\frac{1}{12} b D_{app})$, where we use simple notation D_{app} instead of $D^{(2)}_{app}$ as defined in Section 2.2. The selected three models were constructed by the DKI model when K_{app} hits the upper bound, half and 1/6 the upper bound, respectively. In this way, the ground truth (GT) of K_{app}, θ_W can then be calculated analytically. The anisotropic (positive) diffusion tensors were randomly chosen from a public access data resource `http://academicdepartments.musc.edu/cbi/dki/dke.html`. The last accessible date is 31.12.2018.

We use Philips 32 directions as the gradient scheme, and chose six b values (knots): 62, 249, 560, 996, 1556, 2240 s/mm^2 which were partially from the one that we used to acquire real datasets on human brain and are in an appropriate range with DKI. The noise-free signals were generated from the chosen models (Model 1, 2 or 3) with all the 32 gradient directions isotropically distributed over each shell determined by the different b knot. Hence, the dimension of the design matrix Z_D is 192×6. We generated the Rician noise MR data by corrupting the noise-free signals with three different noise levels: $20, 100$ and 400 that are reasonable when collecting the human brain data, see [36]. For comparison, we fixed the non-attenuation diffusion to be $S_0 = 800$, and used a simple formula for signal to noise ratio, that is, SNR=S_0/σ for $b \sim 0 s/mm^2$ images. Hence three SNR = $2, 8$ and 40 were studied in the experiments.

We first randomly picked up one anisotropic tensor and used Model 1 to generate synthetic data shown in Fig 1. In order to show our results are not occasional, we did further implementation: We chose 100 tensors from the public resource and simulated data with Model 1, 2 and 3, respectively. The diffusivity profile of the selected tensor and the 2D field of the 100 tensors are shown in Fig. 2.

6.2. Simulation 2

The design of this experiment is similar to the one in [15], where the signals were simulated from the biexponential model, see Appendix E, Table A. Data contain 180 voxels from six regions of interest (ROI) which are of different types: gray matter next to cerebration fluid (GM/CSF), gray matter next to white matter (GM/WM), thalamus (TH), putamen and globus pallidus (PU/GP), internal capsule white matter (ICWM), frontal white matter (FWM). Each ROI contains 30 voxels in total. Three shells of

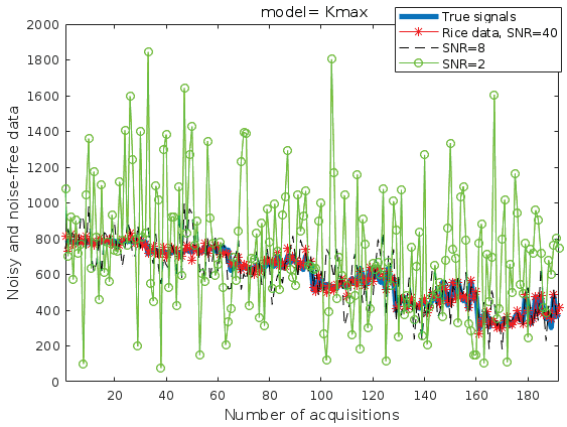
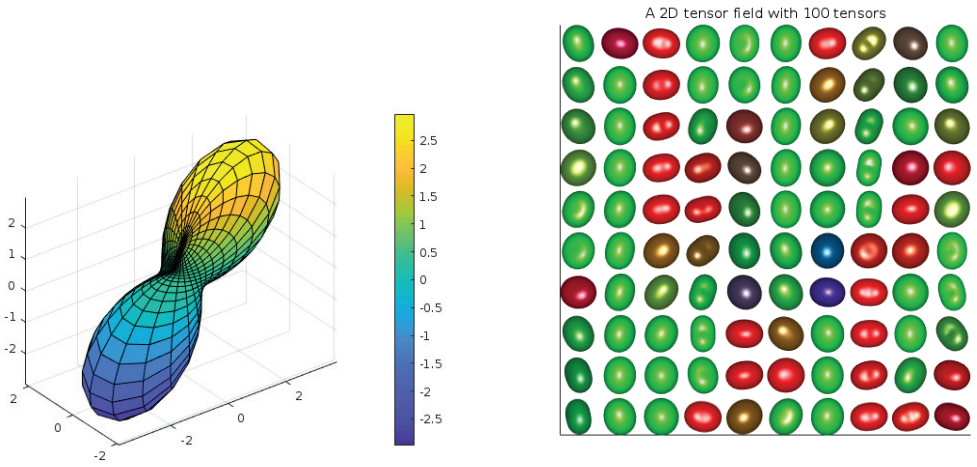


Figure 1: Data were generated by Model 1. The thick blue line describes the noise-free signals (GT), the green circle, dark dash and red star lines represent the Rician noise data with SNR being $2, 8$ and 40, respectively.

b values = 500, 1000, 1500 s/mm^2 and 18 distinct gradients computed by the electrostatic energy minimization algorithm (see Appendix E Table B) were used in the simulation of data. The b knots were increased every 18 gradient directions, hence, for each voxel we have 54 acquisitions. The ground truth (GT) was calculated by the least squares method from the noise-free signals. For comparison, SNRs were chosen within the range [8,40] and were monotonically increasing after every 20 voxels. In the simulation, we fixed $S_0 = 1$. The Rician distributed data were attained by corrupting the signals with noise corresponding to the SNR of the design.

6.3. Real data

We have one subject of a normal volunteer in the case study. The normal brain dataset consists of 2204 diffusion MR-images of the brain, in the form of four $4mm$-thick consecutive axial slices measured by a Philips Achieva 3.0 Tesla MR-scanner. The distance between adjacent slices is $5mm$, and TE/TR is $100ms/25083ms$. The image resolution is 128×128 pixels of size 1.875×1.875 mm^2. In the protocol, we used all the combinations of the 32 gradient directions with the b value varying in the range 0, 62, 249, 560, 996, 1556, 2240 s/mm^2, and the gradients are equally distributed on each shell with 3 repetitions. After masking out the skull and the ventricles, we remain with a region of interest (ROI) containing 18764 voxels to be analyzed.



a) An anisotropic tensor profile

b) The 2D tensor field

Figure 2: Fig. 2a shows a tensor profile drawn on a unit shpere where the color represents the location in the sphere. Fig. 2b depicts a 2D tensor field and was plotted by using the MATLAB fanDTasia ToolBox [37, 38].

7. Results

Next we compare the proposed new methods, EM-IP, CWLS-IP and CWLS-LLS with the completing alternatives, WLS, CWLS-SQP and MLE-SQP.

7.1. Simulation 1

Fig. 3 shows the mean square error (MSE) of the estimated signal decay, $\widehat{S}_{decay} = \exp(Z_D\widehat{\theta}_D + Z_w\widehat{\theta}_w)$, as a function of the b value averaged over the 32 gradient directions by six different methods: WLS, CWLS-LLS, EM-IP, MEL-SQP, CWLS-SQP and CWLS-IP from one randomly selected voxel, where we fixed S_0 for comparison, and data were generated on basis of Model 1. When applying MLE-SQP, CWLS-SQP, CWLS-IP and EP-IP, we used the estimated values of Θ and σ^2 by CWLS-LLS as initials. We use consistent lines to describe different methods: WLS and CWLS-LLS are described by the red cross and the thick dark cross lines, the cyan diamond and the magenta aster lines indicate the results by EM-IP and MLE-SQP, and the thick green dashed circle and black circle lines show MSE of signal decay using CWLS-SQP and CWLS-IP, respectively. According to Fig. 3a, all the methods perform very well when

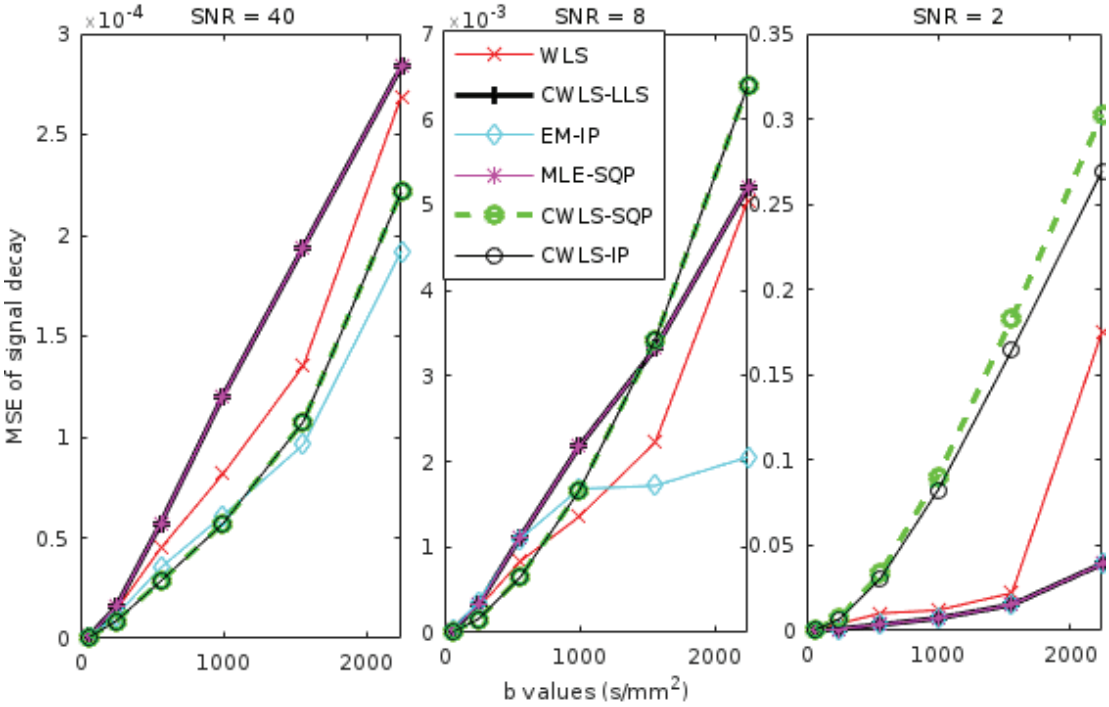


Figure 3: MSE of estimated signal decay ($\widehat{S}_{decay} = \exp(Z_D\widehat{\theta}_D + Z_w\widehat{\theta}_w)$) as a function of the b value from a randomly selected voxel averaged over the 32 gradient directions for every shell. $\widehat{S}_{decay}$ by WLS and CWLS-LLS were computed according to Eq. (4) and are described by the red cross and the thick dark cross lines, respectively. We computed $\widehat{S}_{decay}$ by the other methods through Eq. (7), and use the cyan diamond and the magenta aster lines to indicate the results by EM-IP and MLE-SQP, and the thick green dashed circle and black circle lines to show MSE of $\widehat{S}_{decay}$ using CWLS-SQP and CWLS-IP, respectively.

SNR is high. The results by EM-IP shows outstanding performance when the b value is increasing and are more stable and closer to GP than the alternatives in the cases of three different noise levels. The results by CWLS-LLS are as good as that by MLE-SQP from this experiment for all three different SNRs. When SNR$=2$, the estimates by WLS, CWLS-SQP, CWLS-IP indicate a large deviation from GT, whereas the results by MLE-SQP, CWLS-LLS and EP-IP are equally good.

We did more experiments where we used different synthetic data as described in Section 6.1 to examine 100 voxels, and extended SNR down to 1 as we are more interested in data retrieved in the low regime of SNR. Fig. 4 described MSE of $\widehat{S}_{decay}$ again as a function of the b value averaged over the 32 gradient directions for each shell but in average of 100 voxels, and we assume the voxels are independent of each other. For

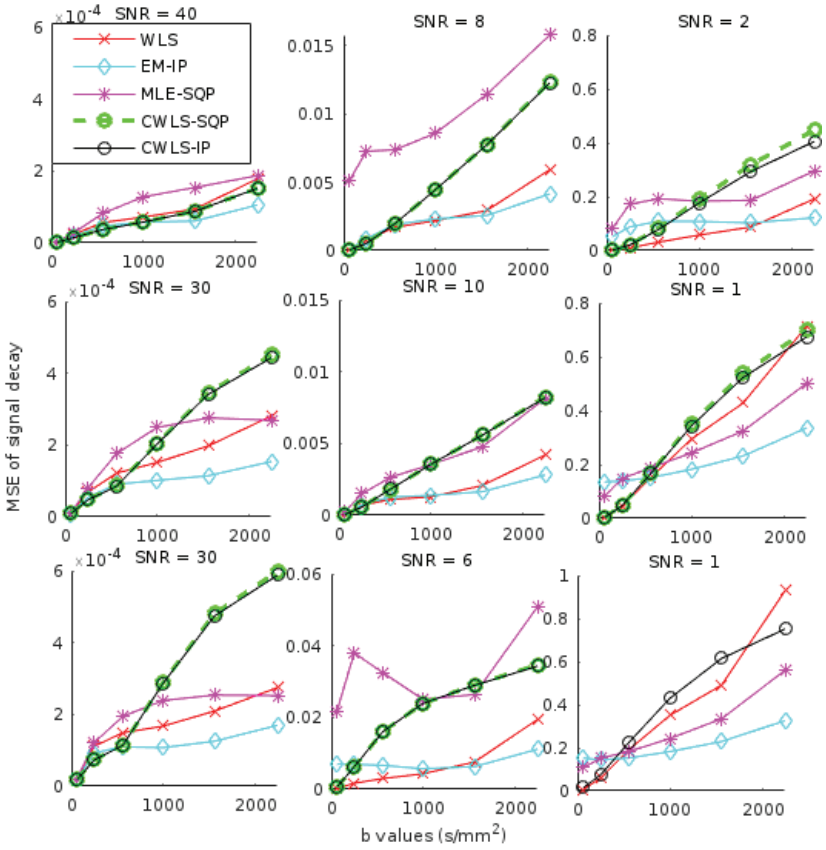


Figure 4: MSE of estimated signal decay ($\widehat{S}_{decay} = \exp(Z_D\widehat{\theta}_D + Z_w\widehat{\theta}_w)$) as a function of the b value in average of 100 randomly selected voxels. For each voxel, $\widehat{S}_{decay}$ were obtained by averaging over all the 32 gradient directions per shell. $\widehat{S}_{decay}$ by WLS were computed according to Eq. (4) and the result is described by the red cross lines, respectively. We computed $\widehat{S}_{decay}$ by the other methods through Eq. (7), and use the cyan diamond and the magenta aster lines to indicate the results by EM-IP and MLE-SQP, and the thick green dashed circle and black circle lines to show MSE using CWLS-SQP and CWLS-IP, respectively.

comparison, the initial values of the methods: MLE-SQP, CWLS-SQP, CWLS-IP and EP-IP were obtained from the results by WLS. Though CWLS-LLS can provide less biased estimates as the case pointed out in Fig. 3, it is unstable in some cases. Hence, we did not show the results by CWL-LLS, as some of them are beyond the regions of the illustrations in Fig. 4. Since the CWLS methods are not based on the correct noise model, they may yield unreliable results, especially when SNR remains in a low regime, which is in agreement with Fig. 4. Also CWLS-SQP (the black circle lines) and CWLS-IP (the thick green dashed circle lines) have roughly similar performance with the first eight datasets, and CWLS-IP shows slightly better achievement, when the

Table 1: Comparison of computational time of EM-IP, MLE-SQP, CWLS-IP, CWLS-SQP.

Simulation 1/ DATA2	EM-IP	MLE-SQP	CWLS-IP	CWLS-SQP
mean	3.5185	2.1462	31.6826	9.2253
max	103.9635	80.5856	105.4403	90.7987
min	0.0625	1.0897	5.8607	0.8128

values of SNR are low. CWLS-SQP gave unreliable results, which is the case in the ninth subplot of Fig. 4. For all the nine illustrations, EM-IP has given the best and stable performance among the alternatives.

In order to compare the computational efficiency, in this experiment we applied a reduced EM algorithm of the proposed method, EM-IP, that we simultaneously update Θ and $\langle \cos\varphi_j \rangle$, though $\langle \cos\varphi_j \rangle$ is thought as a known term from E-step when calculating the score and Hessian matrices, see details in Appendix A, and then we correct the noise level σ by Eq. (14). In such a way, the nonlinear optimization with constraints regarding to optimize Θ only needs run once. We monitor the running time (RT) on average (per voxel), minimum, mean and maximum in seconds for the nine datasets in average, and each dataset contains 100 voxels. The records are listed in Table 1 and was obtained from Cluster, running in parallel in the MATLAB 2018a environment. Table 1 records for all the nine examined datasets, the computation can be done within 2 minutes for all the voxels with different SNR by four listed methods. With Constrained Fisher scoring described in Section 4.1, computation can be more expensive when using the data that simulated with low SNR, but the results are more reliable and stable. In average, MLE-SQP is more efficient and typically yielded better estimates when the b value is high and when SNR is low, but it can be unstable at a few number of voxels, resulting in larger bias than the results by the other alternatives, especially from the data with the low b value and in the high regime of SNR as shown in Fig. 4. In average, EM-IP is the second efficient method among the four illustrations.

7.2. Simulation 2

We plot the mean square error (MSE) of the DKI-derived scalar metrics and compare the performance of the six methods shown in Fig. 5. Figure 5a, 5b, 5c and 5d

depict MSE for MD, DT, MK, KT, respectively, from CWLS-LLS (the thick black cross lines), CWLS-SQP (the thick green dashed circle), CWLS-IP (the black circle lines), MLE-SQP (the magenta aster lines) and EM-IP (the cyan diamond lines). All the subfigures reveal a small precision between the estimates and GT over the selected region of SNR. The precision of all the methods tend to be very small for the results of MD ($\times 10^{-7}$) and DT ($\times 10^{-6}$) as shown in Figure 5a and 5b. We did not show the results by WLS as it does not consider any constraints in DKI. Instead, we describes the results by CWLS-LLS, CWLS-SQP and CWLS-IP that include the three constraints in DKI, despite the fact that they are all based on the log-normal noise model. MLE-SQP provides less precisioned estimates in the first two subfigures (*a* and *b*), because it uses the constraints and the accurate noise model. Our proposed EM method (EM-IP) gives the best performance among the five methods in all the listed metrics. We did not show the results with CWLS-LLS in Figure 5c and 5d as they are beyond the region. Moreover, MSE for MK and KT using MLE-SQP, described in these two figures, describe larger precisioned estimates in some cases of SNR than that of CWLS-IP and CWLS-SQP. The results obtained using CWLS-IP show slight less precision than that by CWLS-SQP. Overall, all the listed methods perform quite well, especially in the high regime of SNR, and our proposed method EM-IP leads to very stable and precise estimates over the selected region of SNR.

Furthermore, we recorded the mean of the percentage of constraints violation at each value of SNR which is described in Fig. 6. It reveals that when SNR increases, the percentage of violation decreases for constraints # 2 and # 3. We got no violation for constraint # 1 in this dataset which imply that all the diffusion tensors estimated using WLS are positive.

7.3. Real data

We first show the anatomic information of the subject of human brain through the metric maps of MD in Fig. 7, FA in Fig. 8 as well as MK in Fig. 9 estimated by the proposed methods, CWLS-IP and EM-IP, respectively. In each of the three figures, the subplots of the first lines depict the results by CWLS-IP, and the results shown on the second lines were obtained by EM-IP. Fig. 7 points out that the results by CWLS-IP are

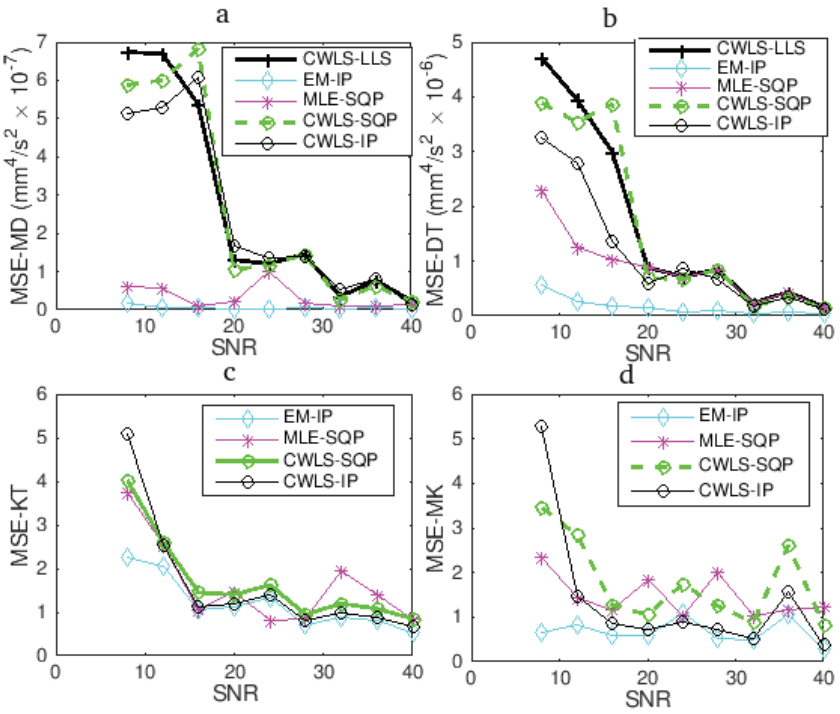


Figure 5: Mean square error (MSE) of mean diffusivity (MD, Fig. 5a), diffusion tensors (DT, Fig. 5b), mean kurtosis (MK, Fig. 5c) and kurtosis tensors (KT, Fig. 5d). The cyan diamond and the magenta aster lines indicate the results by EM-IP and MLE-SQP, and the thick green dashed circle and black circle lines show MSE using CWLS-SQP and CWLS-IP, respectively.

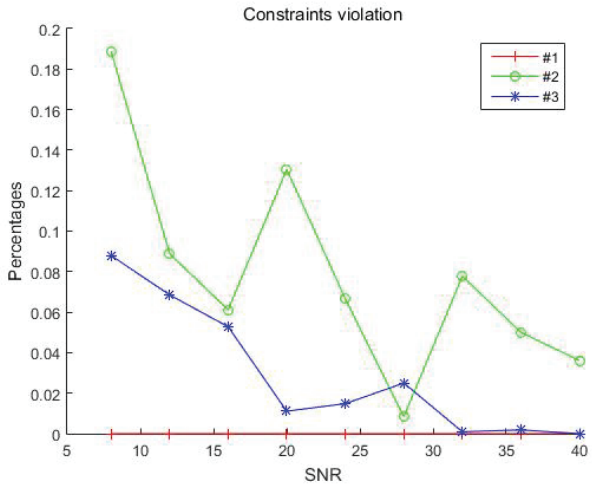


Figure 6: The constraints violation as a function of SNR. The red cross, the blue star and the green circle lines record the percentage of violation of constraint # 1, # 2, # 3, respectively. The percentage of violation for constraints # 2 and # 3 decreases in the increase of SNR.

higher than that by EM-IP from each ROI (slice). The results of voxels in slice 2 shows large differences by these two methods. The estimated values of FA are between (0,1) as described in Fig. 8 by both methods, and the images of FA maps by EM-IP provides more clear visualization of the structural information of ROIs. After comparison at the same scales, the image contrasts by EM-IP represent more clearly the structural information of the brain than those achieved by means of CWLS-IP, especially in Fig. 8 and Fig. 9. Moreover, we also computed mean and standard deviation (std) of MD, FA and MK by the proposed method EM-IP and listed the results in Table 2.

We also monitored constraint violations (CV) and show the spatial layout of CV on the FA maps as in Fig. 11, where CV is evaluated in all 32 encoding directions. As an illustration, we show the results from slice 1 and slice 2, which contain 4819 and 4719 voxels, respectively. The intensities of CV # 1, # 2 and # 3 are on both slices in the range of $[0,16]$ $[0,32]$ and $[0,7]$, respectively. The percentages of CV # 1, # 2 and # 3 in voxels are 0.35% (16/4537), 19.53% (886/4537), 1/4537 and 0.34% (16/4719), 19.37% (914/4719), 1/4719 on slice 1 and on slice 2, respectively.

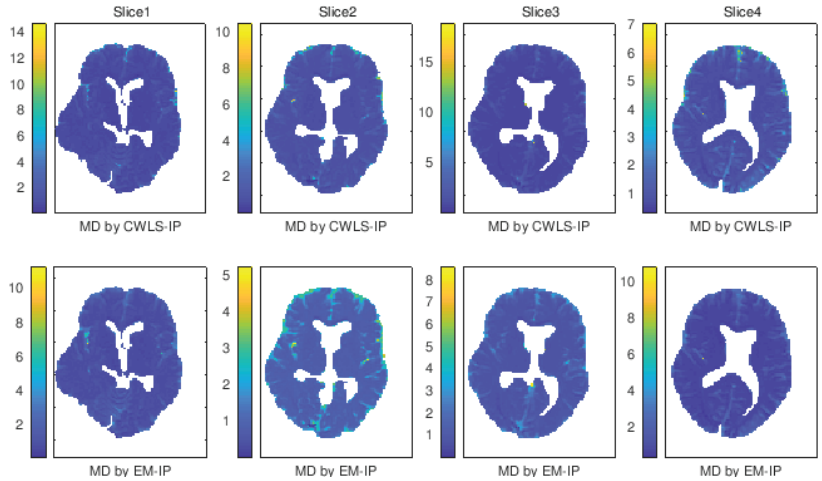


Figure 7: The metric maps of MD obtained by CWLS-IP and EM-IP from four consecutive slices of human brain. The values of estimated MD are between (0,18) $\times 10^{-3} mm^2/s$ by CWL-IP and are between (0,10) $\times 10^{-3} mm^2/s$ by EM-IP for the same ROIs, respectively.

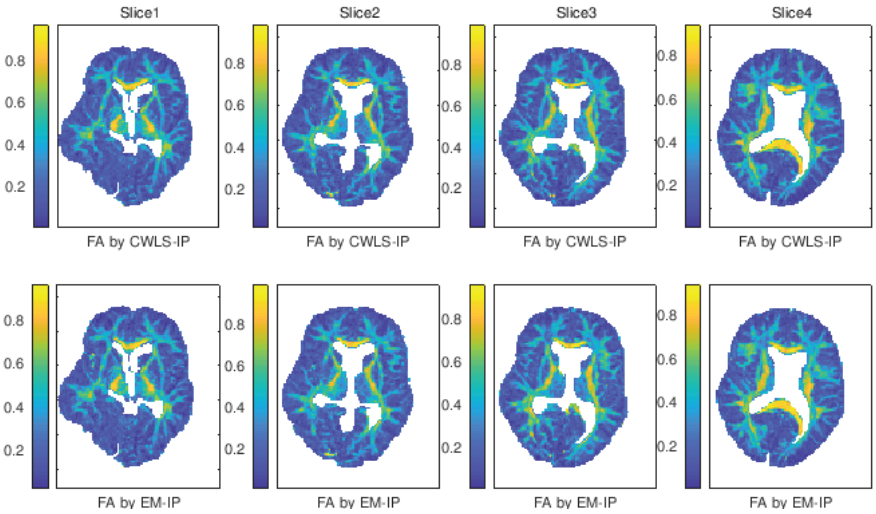


Figure 8: The metric maps of FA obtained using CWLS-IP and EM-IP from four consecutive slices of human brain. The values of estimated FA are between (0,1) by both methods.

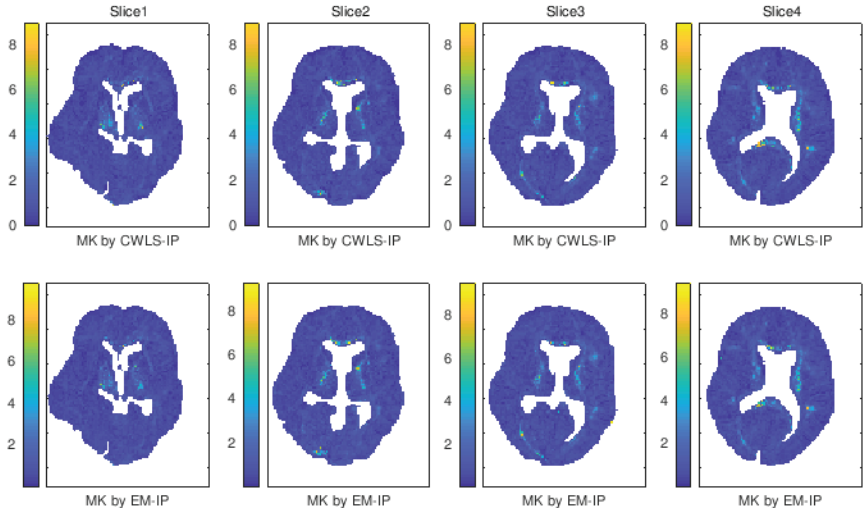


Figure 9: The metric maps of MK obtained using CWLS-IP and EM-IP from four consecutive slices of human brain. The values of estimated MK are between (0,8). The subfigures of the first lines are with CWLS-IP, and the plots of the second line were obtained by EM-IP.

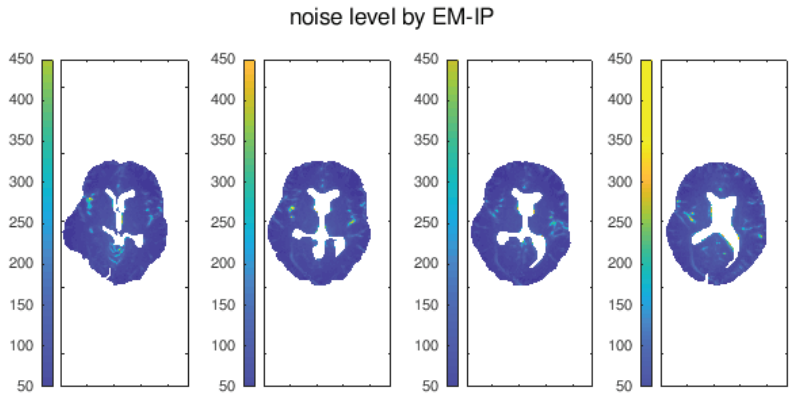


Figure 10: The estimated noise level ($\hat{\sigma}$) by EM-IP. The values are in the region between 50 and 450 for four ROIs from the subject of human brain.

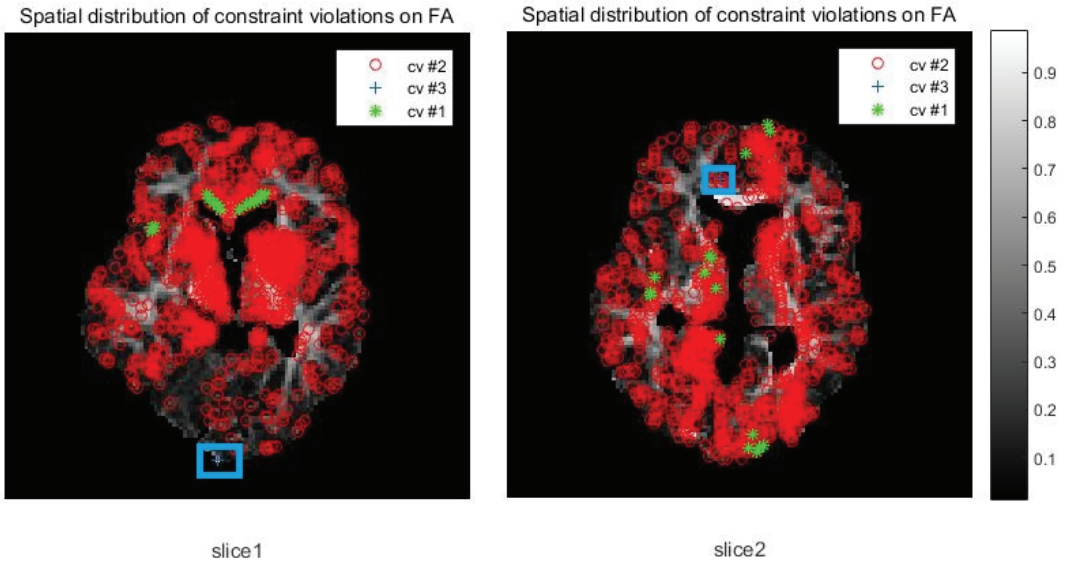


Figure 11: Spatial distributions of constraint violations (CV) on the FA maps of two illustrated slices. The red circles show CV # 2 and the green stars depict CV # 1. We mark CV # 3 by a blue rectangle as it has very low percentage on these two slices.

Table 2: Mean and standard deviation (std) of MD, FA and MK by EM-IP.

real data	Slice 1		Slice 2	
	mean	std	mean	std
MD	1.5548×10^{-3}	3.063×10^{-3}	1.0479×10^{-3}	5.184×10^{-4}
FA	0.2553	0.1737	0.2647	0.1817
MK	0.6855	0.4504	0.7401	0.5736
real data	Slice 3		Slice 4	
	mean	std	mean	std
MD	1.0856×10^{-3}	5.649×10^{-4}	1.0771×10^{-3}	5.337×10^{-4}
FA	0.2753	0.1860	0.2844	0.2081
MK	0.7866	0.5884	0.7951	0.6735

8. Discussion

The contributions of the Fisher scoring algorithm with the correct forms of the Hessian matrices lead to much reliable results without any increase of computational burden being in agreement with the results produced by CWLS-IP and EM-IP with comparison of MLE-SQP and CWLS-SQP, respectively. Especially in our case, θ_D is a function of L, which provides a possibility to calculate the essential Fisher information for updating the parameter L in the Fisher scoring method. Compared with the observed information (or so-called empirical Fisher information) $\mathscr{J}(L)$, the algebraically simple formula of the Fisher information will lead to substantially less demanding computation. The stability of the algorithm can be achieved by means of proper regularization.

In terms of computation efficiency, in the simulation studies we introduced a "reduced" EM algorithm that we update Θ and $\langle\cos\varphi_j\rangle$ simultaneously though $\langle\cos\varphi_j\rangle$ is considered as a known term from E-step. In such a way, the constrained nonlinear optimization can only run once without increasing computational burden, and the noise can be corrected afterwards. There are only a few works, for instance [18, 36] with studies on the noise in diffusion MRI. The Von Mises data augmentation described in this paper provides an alternatively efficient way to correct the noise in diffusion MRI.

In the study of real data, we depict the scalar metric maps of MD, FA and MK,

which are commonly used to explore the anatomic information of the brain. MD describes the diffusion anisotropy in terms of diffusivity, FA measures the diffusion anisotropy of water molecules at each voxel, and MK describes the degree of discrepancy of the diffusion displacement distribution from the Gaussian distribution. These maps together can detect nearly all types of fibers in the white matter and give detailed structural information of brain anatomy.

The proposed CWLS-LLS are sometimes surprisingly as good as other methods that considered all the constraints in DKI, this is probably because in the iteration step of CWLS-LLS, Step 3 breaks the Gaussian assumption and forces the estimates fit the Rician model in Eq. (2) as close as possible. The method has much less computation burden compared with others. However, it is unstable in general, hence, further investigation and improvements are needed.

This work has some limitations that should be taken into account in the future work: 1) The number of cases in real data was small and only from a healthy volunteer. 2) We did not analyze inside individual brain regions in more detail but rather focus on statistical analysis of the methods proposed in this work. 3) In this work, we assumed each voxel is independent of the other, hence correlations between spatial neighborhood of the voxels have not been considered in the estimation. For this limitation, we have proposed a general method, which can be extended to this work, see [39].

9. Conclusion

In this work, we proposed several methods: EM-IP, CWLS-IP and CWLS-LLS for the estimation with constrained DKI diffusion MRI. EM-IP has shown better performance in our studies compared with the alternatives and other popular methods. Using the state-of-the-art statistical methodology of data augmentation, we are able to work with a generalized linear model (GLM) of the joint likelihood derived from the Rician density. The positivity constraints are imposed by the Cholesky decomposition and the new parametrization of the ternary quadratic (TQ) of rank 2 and kurtosis tensors, respectively. The whole scheme is not only for updating the tensor parameters simultaneously but also for updating the noise parameters and the unattenuated signal. To

apply this scheme for other simpler models such as DTI, the multi-tensor model or other DWI alternatives are straightforward. The new augmentation method leads to less biased estimators, which was expected due to theoretical reasoning and had been supported by the experiments.

10. Acknowledgement

The author specially thanks Doctor Dario Gasbarra and Juha Railavo for collecting real data and obtaining permissions, and Doctor Salme Kärkkäinen for proofreading the manuscript. The author is grateful to Emeritus Professor Antti Penttinen from the University of Jyväskylä who carefully read and insightfully commented on the manuscript multiple times from the draft to the revised versions, in which this work can be finally represented. Moreover, the author is grateful to the CSC-IT Center for Science Ltd. for technique support. This work was funded by Doctoral Program in Computing and Mathematical Sciences (COMAS) and Department of Mathematics and Statistics, University of Jyväskylä. No potential conflict of interest was reported by the author.

Appendix

Appendix A. Fisher information of L and θ_Q, and the barrier method

The Fisher information is

$$\mathscr{I}(L^{(k)},\lambda^{(k)}) = \mathbb{E}\big[-H(L^{(k)},\lambda^{(k)})\big] = -\mathbb{E}\big[\bigtriangledown^2 f(L) + \sum_{j=1}^{m} \lambda_j \bigtriangledown^2 g_j(L)\big]$$

$$= -\left[J_L^T(\bigtriangledown^2 f(\theta_D))J_L + \bigtriangledown f(\theta_D)\frac{\partial^2 \theta_D(L)}{\partial L_k \partial L_h}\right] - \sum_{j=1}^{m} \lambda_j M_{D_j}$$

$$= -(\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ J_L^T \left(2(S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 Z_{Dj}^T Z_{Dj} - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} Z_{Dj}^T Z_{Dj} \right) J_L \right\} - \sum_{j=1}^{m} \lambda_j M_{D_j},$$

where λ is the Lagrangian multiplier,

$$M_{D_j} := \bigtriangledown^2 g_j(L)$$

$$= \left[Z_{Dj} \frac{\partial^2 \theta_D(L)}{\partial L_k \partial L_h} \right] = \begin{pmatrix} 2Z_{D_{1j}} & & & Z_{D_{4j}} & Z_{D_{5j}} & \\ & 2Z_{D_{2j}} & & & & Z_{D_{6j}} \\ & & 2Z_{D_{3j}} & & & \\ & Z_{D_{4j}} & & 2Z_{D_{2j}} & & Z_{D_{6j}} \\ Z_{D_{5j}} & Z_{D_{6j}} & & & 2Z_{D_{3j}} & \\ & & & Z_{D_{6j}} & & 2Z_{D_{3j}} \end{pmatrix},$$

and $[-H(L,\lambda)]$ is known as the empirical information. The gradient $\bigtriangledown f(L) \in \mathbb{R}^d$ of $f(L)$ at the current recursion is

$$\bigtriangledown f(L^{(k)}) = (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ (S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 J_L Z_{D_j}^T - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} J_L Z_{D_j}^T \right\}$$

and

$$\bigtriangledown g(L^{(k)}) = Z_{Dj} J_L^{(k)}.$$

In order to ease the computation in the update of θ_Q, we simply compute the empirical Fisher information

$$\mathscr{I}(\theta_Q^{(k)}, \lambda^{(k)}) = -H(\theta_Q^{(k)}, \lambda^{(k)}) = -\bigtriangledown^2 f(\theta_Q) - \sum_{j=1}^{m} \lambda_j \bigtriangledown^2 g_j(\theta_Q)$$

$$= -(\sigma^{-2})^{(k)} \sum_{j=1}^{m} \Big\{ 8(S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 \theta_Q^T P_j^T P_j \theta_Q + 2(S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 P_j$$

$$- 4S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} \theta_Q^T P_j^T P_j \theta_Q - 2S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} P_j \Big\} - \sum_{j=1}^{m} 4\lambda_j P_j,$$

where the gradient of $f(\theta_Q)$ is

$$\bigtriangledown f(\theta_Q) = (\sigma^{-2})^k \sum_{j=1}^{m} \left\{ 2(S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 P_j \theta_Q - 2S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} P_j \theta_Q \right\},$$

and

$$\bigtriangledown g(\theta_Q) = 4(\theta_Q^T)^{(k)} P_j.$$

The barrier method [19, 40], also known as the primal-dual *interior point* method (IP), is among a few successful algorithms in solving such a kind of optimization prob-

lems, where the inequality constraints are imposed by a barrier ν, so that

$$g_j(\Theta) - \nu_j = 0, \qquad j = 1, \cdots, m. \qquad \nu_j \geq 0.$$

The scheme for solving the parameters Θ, λ and ν is then

$$\mathbb{S} J_{ac}^{(-1)}, \tag{.1}$$

where $\mathbb{S}$ is the matrix containing the score of Θ, λ and ν, and J_{ac} is a Jacobian matrix given by

$$J_{ac}(\Theta, \lambda, \nu) = \begin{pmatrix} \mathscr{I}(\Theta, \lambda) & 0 & A(\Theta)^\top \\ 0 & diag(\lambda) & diag(\nu) \\ A(\Theta) & \mathbb{I}_{m\times 1} & 0 \end{pmatrix},$$

where $diag(\cdot)$ is a diagonalizing operator to construct the vector to be a $m \times m$ matrix, and $A(\theta) := \bigtriangledown g(\theta)$ is a $d \times m$ matrix.

Appendix B. Calculation of $< \cos \varphi_j >$

By the moment generating function of the Von Mises distribution, we have

$$E_{S,\sigma^2}\big(\exp(\lambda \cos(\varphi))\big|Y\big) = \frac{I_0(\lambda + YS/\sigma^2)}{I_0(YS/\sigma^2)}, \quad \Longrightarrow$$
$$E_{S,\sigma^2}\big(\cos(\varphi)\big|Y\big) = \frac{\partial}{\partial \lambda} \frac{I_0(\lambda + YS/\sigma^2)}{I_0(YS/\sigma^2)}\bigg|_{\lambda=0} = \frac{\partial}{\partial z} \log I_0(z)\bigg|_{z=YS/\sigma^2} = \frac{I_1(YS/\sigma^2)}{I_0(YS/\sigma^2)} \quad ,$$

where S is the signal and $I_k(z)$ is the modified Bessel function of first kind. This gives

$$\big\langle \cos \varphi_j \big\rangle = \frac{I_1\big(Y_j S_0 \exp(Z_j \theta)\sigma^{-2}\big)}{I_0\big(Y_j S_0 \exp(Z_j \theta)\sigma^{-2}\big)}.$$

Appendix C. Fisher scoring method for L

Let's $\zeta_j^{(k)} = \exp(Z_{Dj}\theta_D^{(k)}), \psi_j^{(k)} = \exp\Big((\theta_Q^{(k)})^T P_j \theta_Q^{(k)}\Big)$ and $\tau_j^{(k)} = Y_j \big\langle \cos(\varphi_j) \big\rangle^{(k)}$. The score of θ_D is the first derivative of Eq. (15) w.r.t. θ_D given as

$$\bigtriangledown q(\theta_D) = (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \Big\{ (S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 Z_D^T - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} Z_D^T \Big\}, \tag{.1}$$

and the Hessian matrix are

$$\bigtriangledown^2 q(\theta_D) = (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ 2(S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 Z_D^T Z_D - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} Z_D^T Z_D \right\}, \tag{.2}$$

and the observed information $(\theta_D) = -\bigtriangledown^2 q(\theta_D)$ is defined as the Hessian multiplied by -1.

The score of L expresses

$$\bigtriangledown q(L) = (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ (S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 Z_D^T J_L - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} Z_D^T J_L \right\}, \tag{.3}$$

and the corresponding Hessian matrix is

$$\bigtriangledown^2 q(L) = J_L^T (\bigtriangledown^2 q(\theta)) J_L + \bigtriangledown q(\theta_D) \frac{\partial^2 \theta_D(L)}{\partial L_k \partial L_h} \tag{.4}$$

$$= (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ J_L^T \left(2(S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 Z_{Dj}^T Z_{Dj} - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} Z_{Dj}^T Z_{Dj} \right) J_L \right\}$$

$$- (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ \left((S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} \right) M_j \right\}, \tag{.5}$$

where

$$M_j = Z_{Dj} \frac{\partial^2 \theta_D(L)}{\partial L_k \partial L_h} = \begin{pmatrix} 2Z_{1j} & & & Z_{4j} & Z_{5j} & \\ & 2Z_{2j} & & & & Z_{6j} \\ & & 2Z_{3j} & & & \\ & Z_{4j} & & 2Z_{2j} & & Z_{6j} \\ Z_{5j} & Z_{6j} & & & 2Z_{3j} & \\ & & & Z_{6j} & & 2Z_{3j} \end{pmatrix}.$$

The Fisher information is given by

$$\left\langle \mathscr{I}(L)^{(k)} \right\rangle := \mathbb{E}\left[-\bigtriangledown^2 \log \pi(y; \theta_D(L)) \right] =$$

$$- (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ J_L^T \left(2(S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 Z_{Dj}^T Z_{Dj} - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} Z_{Dj}^T Z_{Dj} \right) J_L \right\},$$

with the expectation at $\tilde{\theta}_D$, the current value of θ_D,

$$\mathbb{E}\left[\bigtriangledown q(\theta_D) \right] = 0 \qquad \text{and}$$

$$\mathbb{E}\left[\bigtriangledown^2 q(\theta_D) \right] = (\sigma^{-2})^{(k)} \sum_{j=1}^{m} \left\{ (S_0^{(k)})^2 (\zeta_j^{(k)})^2 (\psi_j^{(k)})^2 Z_D^T - S_0^{(k)} \tau_j^{(k)} \zeta_j^{(k)} \psi_j^{(k)} Z_D^T \right\}.$$

Note that θ_D is a function of L which provides the possibility to calculate the essential Fisher information which is the expected value of (or minus) Hessian matrix for updating L in the Fisher scoring method. Compared with the observed information $\mathscr{I}(L)$, the algebraically simpler formula of the Fisher information will substantially reduce computation, and the algorithm is much more stable regarding the singularity than the observed information matrix. Details can be found in [41].

Appendix D. The Gram matrix

We extract $\widehat{\theta}_W$ from the Gram matrix [42] computed by $G^* = \widehat{Q}^T\widehat{Q}$, that given by

$$G^* = \begin{pmatrix} \theta_W(1) & a & b & \theta_W(10) & \theta_W(11) & 2d \\ a & \theta_W(2) & c & \theta_W(12) & 2e & \theta_W(13) \\ b & c & \theta_W(3) & 2f & \theta_W(14) & \theta_W(15) \\ \theta_W(10) & \theta_W(12) & 2f & 4\theta_W(4)-2a & \theta_W(7)-2d & \theta_W(8)-2e \\ \theta_W(11) & 2e & \theta_W(14) & \theta_W(7)-2d & 4\theta_W(5)-2b & \theta_W(9)-2f \\ 2d & \theta_W(13) & \theta_W(15) & \theta_W(8)-2e & \theta_W(9)-2f & 4\theta_W(6)-2c \end{pmatrix},$$

where $\widehat{Q}$ is a 6×3 matrix constructing by all the elements in $\widehat{\theta}_Q$,
and $\theta_W(1) = G^*(1,1), \theta_W(2) = G^*(2,2), \theta_W(3) = G^*(3,3), \theta_W(4) = 1/4G^*(4,4)$
$+ 1/2G^*(1,2), \theta_W(5) = 1/4G^*(5,5) + 1/2G^*(1,3), \theta_W(6) = 1/4G^*(6,6)$
$+ 1/2G^*(2,3), \theta_W(7) = G^*(4,5) + 2G^*(1,6), \theta_W(8) = G^*(4,6) + 2G^*(2,5),$
$\theta_W(9) = G^*(5,6) + 2G^*(3,4), \theta_W(10) = G^*(4,1), \theta_W(11) = G^*(5,1),$
$\theta_W(12) = G^*(4,2), \theta_W(13) = G^*(6,2), \theta_W(14) = G^*(5,3), \theta_W(15) = G^*(6,3).$

Appendix E.

The values of these six ROIs were taken from [43] with the biexponential diffusion model.

This set of gradient directions was taken from [44], point set 1, which was computed by electrostatic energy minimization algorithm and shown the advantage of maintaining the optimal cover.

Table E.3: Parameters from normal human brains

ROI	$D_{in}[mm^2/s \times 10^{-3}]$	$D_{ex}[mm^2/s \times 10^{-3}]$	f_{in}
GM/CSF	1.479 ± 0.166	0.466 ± 0.017	0.490 ± 0.012
GM/WM	1.142 ± 0.106	0.338 ± 0.027	0.622 ± 0.038
TH	1.320 ± 0.164	0.271 ± 0.040	0.617 ± 0.069
PU/GP	1.609 ± 0.039	0.257 ± 0.026	0.648 ± 0.028
FWM	1.155 ± 0.046	0.125 ± 0.026	0.648 ± 0.050
ICWM	1.215 ± 0.024	0.183 ± 0.009	0.637 ± 0.020

References

[1] P. J. Basser, J. Mattiello, D. Le Bihan, Estimation of the effective self-diffusion tensor from the NMR spin echo, Journal of Magnetic Resonance, Series B 103 (3) (1994) 247–254.

[2] P. J. Basser, J. Mattiello, R. Turner, D. Le Bihan, Diffusion tensor echo-planar imaging of human brain, in: Proceedings of the SMRM, Vol. 584, 1993.

[3] P. J. Basser, J. Mattiello, D. LeBihan, MR diffusion tensor spectroscopy and imaging., Biophysical Journal 66 (1) (1994) 259.

[4] D. S. Tuch, R. Weisskoff, J. Belliveau, V. Wedeen, High angular resolution diffusion imaging of the human brain, in: Proceedings of the 7th Annual Meeting of ISMRM, Philadelphia, Vol. 321, 1999.

[5] D. S. Tuch, Diffusion MRI of complex tissue structure, Ph.D. thesis, Citeseer (2002).

[6] L. R. Frank, Characterization of anisotropy in high angular resolution diffusion-weighted MRI, Magnetic Resonance in Medicine 47 (6) (2002) 1083–1099.

[7] E. Özarslan, T. H. Mareci, Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging, Magnetic Resonance in Medicine 50 (5) (2003) 955–965.

Table E.4: Optimized 18 gradient directions

0.737068	-0.568030	0.366160
0.795763	0.431108	0.425331
-0.822530	0.367692	0.433874
0.000650	0.985575	0.169239
0.228998	0.150756	0.961682
-0.412439	-0.753502	0.511984
-0.358616	0.232844	0.903979
-0.891249	-0.417614	0.176844
0.319924	-0.498679	0.805586
0.309857	0.667672	0.676907
0.579701	-0.807043	-0.112374
-0.209598	-0.358489	0.909700
0.990653	-0.112342	0.077367
0.153276	-0.903274	0.400754
0.530172	0.845386	0.065124
-0.282930	0.716688	0.637423
0.720077	-0.052737	0.691887
-0.733882	-0.178601	0.655377

[8] J. H. Jensen, J. A. Helpern, A. Ramani, H. Lu, K. Kaczynski, Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging, Magnetic Resonance in Medicine 53 (6) (2005) 1432–1440.

[9] H. Lu, J. H. Jensen, A. Ramani, J. A. Helpern, Three-dimensional characterization of non-Gaussian water diffusion in humans using diffusion kurtosis imaging, NMR in Biomedicine 19 (2) (2006) 236–247.

[10] J. H. Jensen, J. A. Helpern, MRI quantification of non-Gaussian water diffusion by kurtosis analysis, NMR in Biomedicine 23 (7) (2010) 698–710.

[11] L. Qi, D. Han, E. X. Wu, Principal invariants and inherent parameters of diffusion kurtosis tensors, Journal of Mathematical Analysis and Applications 349 (1) (2009) 165–180.

[12] A. J. Steven, J. Zhuo, E. R. Melhem, Diffusion kurtosis imaging: An emerging technique for evaluating the microstructural environment of the brain, American Journal of Roentgenology 202 (1) (2014) W26–W33.

[13] A. Tabesh, J. H. Jensen, B. A. Ardekani, J. A. Helpern, Estimation of tensors and tensor-derived measures in diffusional kurtosis imaging, Magnetic Resonance in Medicine 65 (3) (2011) 823–836.

[14] A. Ghosh, T. Milne, R. Deriche, Constrained diffusion kurtosis imaging using ternary quartics & MLE, Magnetic Resonance in Medicine 71 (4) (2014) 1581–1591.

[15] J. Veraart, W. Van Hecke, J. Sijbers, Constrained maximum likelihood estimation of the diffusion kurtosis tensor using a Rician noise model, Magnetic Resonance in Medicine 66 (3) (2011) 678–686.

[16] H. Lu, J. Jensen, C. Hu, M. Falangola, A. Ramani, S. Ferris, J. Helpern, Alterations in cerebral microstructual integrity in normal aging and in Alzheimers disease: a multi-contrast diffusion MRI study, in: Proc Int Soc Magn Reson Med, Vol. 14, 2006, p. 723.

[17] J. A. Helpern, C. Lo, C. Hu, M. Falangola, O. Rapalino, J. H. Jensen, Diffusional kurtosis imaging in acute human stroke, in: Proceedings 17th Scientific Meeting, International Society for Magnetic Resonance in Medicine, Vol. 17, 2009, p. 3493.

[18] J. Liu, D. Gasbarra, J. Railavo, Fast estimation of diffusion tensors under Rician noise by the EM algorithm, Journal of Neuroscience Methods 257 (2016) 147–158.

[19] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge University Press, 2004.

[20] R. M. Henkelman, Measurement of signal intensities in the presence of noise in MR images, Medical physics 12 (2) (1985) 232–233.

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

[22] C. G. Koay, L.-C. Chang, J. D. Carew, C. Pierpaoli, P. J. Basser, A unifying theoretical and algorithmic framework for least squares methods of estimation in diffusion tensor imaging, Journal of Magnetic Resonance 182 (1) (2006) 115–125.

[23] H. Zhu, H. Zhang, J. G. Ibrahim, B. S. Peterson, Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data, Journal of the American Statistical Association 102 (480) (2007) 1085–1102.

[24] A. Barmpoutis, J. Zhuo, Diffusion kurtosis imaging: Robust estimation from DW-MRI using homogeneous polynomials, in: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, IEEE, 2011, pp. 262–265.

[25] M. A. Tanner, W. H. Wong, The calculation of posterior distributions by data augmentation, Journal of the American Statistical Association 82 (398) (1987) 528–540.

[26] C. Robert, G. Casella, Monte Carlo statistical methods, Springer Science & Business Media, 2013.

[27] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian data analysis, Vol. 2, Taylor & Francis, 2014.

[28] S. Mori, Introduction to diffusion tensor imaging, Elsevier, 2007.

[29] N. I. Fisher, T. Lewis, B. J. Embleton, Statistical analysis of spherical data, Cambridge University Press, 1987.

[30] P. McCullagh, J. A. Nelder, Generalized linear models, Vol. 37, CRC press, 1989.

[31] R. Sundberg, Maximum likelihood theory for incomplete data from an exponential family, Scandinavian Journal of Statistics (1974) 49–58.

[32] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (methodological) (1977) 1–38.

[33] G. McLachlan, T. Krishnan, The EM algorithm and extensions, Vol. 382, John Wiley & Sons, 2007.

[34] J. Fan, Y. Yuan, On the convergence of a new Levenberg-Marquardt method, Thechnical Report, AMSS, Chinese Academy of Sciences (2001) 1–11.

[35] N. Yamashita, M. Fukushima, On the rate of convergence of the Levenberg-Marquardt method, in: Topics in Numerical Analysis, Springer, 2001, pp. 239–249.

[36] M. G. Péreza, A. Concib, A. B. Morenoc, V. H. Andaluza, J. A. Hernándezd, Estimating the Rician noise level in brain MR image, in: Andescon, IEEE, 2014, pp. 1–1.

[37] A. Barmpoutis, M. S. Hwang, D. Howland, J. R. Forder, B. C. Vemuri, Regularized positive-definite fourth order tensor field estimation from DW-MRI, NeuroImage 45 (1) (2009) S153–S162.

[38] A. Barmpoutis, B. C. Vemuri, A unified framework for estimating diffusion tensors of any order with symmetric positive-definite constraints, in: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, IEEE, 2010, pp. 1385–1388.

[39] D. Gasbarra, J. Liu, J. Railavo, Data augmentation in Rician noise model and Bayesian diffusion tensor imaging, 2014.

[40] A. P. Ruszczyński, Nonlinear optimization, Vol. 13, Princeton University Press, 2006.

[41] P. J. Green, Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, Journal of the Royal Statistical Society. Series B (Methodological) (1984) 149–192.

[42] A. Barmpoutis, B. Jian, B. C. Vemuri, T. M. Shepherd, Symmetric positive 4 th order tensors & their estimation from diffusion weighted MRI (2007) 308–319.

[43] S. E. Maier, R. V. Mulkern, Biexponential analysis of diffusion-related signal decay in normal human cortical and deep gray matter, Magnetic Resonance Imaging 26 (7) (2008) 897–904.

[44] P. A. Cook, M. Symms, P. A. Boulby, D. C. Alexander, Optimal acquisition orders of diffusion-weighted MRI measurements, Journal of Magnetic Resonance Imaging 25 (5) (2007) 1051–1058.

IV

VARIATIONAL BAYES ESTIMATION IN CONSTRAINED KURTOSIS DIFFUSION IMAGING UNDER A RICIAN NOISE MODEL

by

Liu, J, Gasbarra, D & Railavo, J. 2019

Submitted manuscript

Variational Bayes Estimation in Constrained Kurtosis Diffusion Imaging under a Rician Noise Model

Jia Liu[a,b,*], Dario Gasbarra[b], Juha Railavo[c]

[a] *Department of Mathematics and Statistics, University of Jyväskylä, P.O. Box 35 (MaD) FI40014 Finland*
[b] *Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68 FI00014 Finland*
[c] *Helsinki University Hospital*

Abstract

The analysis of diffusion MR-data is often based on the simplifying assumption that the diffusion of water molecules follows locally a centered Gaussian distribution. However, diffusional non-Gaussianity is a common scenario in biological tissues due to potential barriers and compartments. Diffusion kurtosis imaging (DKI) is an extension of diffusion tensor imaging (DTI), quantifying both Gaussian and non-Gaussian diffusivity by means of 2nd-order tensors and 4th-order kurtosis tensors, and is considered highly useful in the diagnosis of brain disorders. The model contains three physical constraints on the diffusivity. The correct Rician likelihood is preferred in the estimation, especially for the MRI-data in the regime of low signal to noise ratio, in order to obtain more reliable and accurate estimates.

Diffusion MRI-data are subject to noise and artefacts. A general method for image denoising is image regularization, where the parameter estimators at each voxel depend also on the estimators at its neighbours. Often the regularization step is applied at the second stage after estimating the tensor parameters from data. An alternative approach is to perform the estimation and regularization steps simultaneously.

In this work we propose an original and efficient Bayesian computational method for DKI including an imaging regularization technique. We use Von Mises data augmentation to reduce the computational difficulties from the Rician likelihood. We built a Bayesian model for approximative posterior inference of the DKI parameters and ap-

*Corresponding author
Email address: jia.2.liu@jyu.fi; jia.liu@helsinki.fi (Jia Liu)

plied the variational Bayes (VB) method for posterior computation. A regularization technique is suggested for smoothing the images. The three constraints in DKI are considered in the methodology. Experiments are conducted on synthetic data to evaluate the performance of the proposed method and on real data from a case-control study regarding Lévy body dementia.

Keywords: Data Augmentation, Delta Method, Diffusion Kurtosis Imaging, Gaussian Markov Random Field, Lévy Body Dementia, Non-Gaussian Diffusion, Nonlinear Constraints, Regularization, Rician Likelihood, Tensor Positivity, Variational Bayes, Von Mises Distribution.

---

1. Introduction

Diffusion MRI is capable of measuring the displacement diffusion of water molecules and provides a unique insight into image contrasts reflecting anatomical architectures inside the organic tissue. Diffusion tensor imaging (DTI), introduced by Basser et al. (1993, 1994) is an established noninvasive imaging technique based on the diffusion weighted MR measurements. It extracts the neurostructural information by characterizing the Gaussian diffusion anisotropy through a three-dimensional (3D) 2nd-order tensor Θ at each volume element (voxel). The probability distribution of water diffusion, however, is extremely difficult to model due to the complex microstructure of the underlying biological tissues. Among all the biological tissues, the human brain is the most interesting media, which is rich in microstructures such as cell membranes, boundaries and other complex compartments, and hence the water diffuses in a non-Gaussian way. Tuch et al. (1999); Tuch (2002) proposed the High order Angular Resolution Imaging (HARDI) to overcome the limitations of DTI, that is, in HARDI the probability distribution of water diffusion is not restricted to be jointly Gaussian, and is hence widely used to characterize non-Gaussian diffusion processes.

Among the recent popular imaging protocols, the diffusion kurtosis imaging (DKI) Jensen et al. (2005); Helpern et al. (2009); Jensen and Helpern (2010) is an extension of DTI combining the advantages of HARDI. It attempts to quantify the degree of diffusional non-Gaussianity by introducing a 4th-order tensor (W), using a statistical

metric of "peakedness" named *kurtosis*. The DKI model (Jensen et al., 2005; Jensen and Helpern, 2010) for signal intensity is given by

$$S = S_0 \exp\left(-b\, \Theta(\mathbf{g}) + \frac{b^2\, \overline{tr(\Theta)}^2}{6} W(\mathbf{g}) \right), \tag{1}$$

where S_0 is the baseline signal without diffusion weighting or the so-called baseline signal intensity, b is the factor of diffusion weighted sequences summarizing the impact of the gradient strength on the diffusion weighted images, and $\mathbf{g}$ indicates the gradient direction as a vector on the unit sphere. The diffusion component $\Theta(\mathbf{g})$ and the kurtosis component $W(\mathbf{g})$, are respectively quadratic and quartic forms, symmetric in the $\mathbf{g}$-coordinates.

Since DKI is a gradient based imaging technique in order to identify the parameters Θ and W in Eq. (6), the acquisitions are required at least in 15 different gradient directions with three distinct b-values. Furthermore, to be consistent with the physical relevance of the diffusion phenomenon, the following constraints are imposed, see also Veraart et al. (2011); Ghosh et al. (2014):

- The diffusion tensor Θ should be positive definite (#1).
- The lower (#2) and upper (#3) bounds of $W(\mathbf{g})$ are

$$0 \leq W(\mathbf{g}) \leq \frac{3\Theta(\mathbf{g})}{\overline{tr(\Theta)}^2}, \tag{2}$$

  in order to obtain a signal decreasing with respect to the b-values in the acquisition range.

The positivity constraint for Θ can be taken into account by using the Cholesky parametrization $\Theta = LL^\top$. Several authors have proposed solutions to guarantee higher order tensor positivity: Barmpoutis et al. (2007); Barmpoutis and Zhuo (2011) use a sum of squares of quadratic forms to represent the 4th-order tensor; Qi et al. (2010) propose a positive semi-definite diffusion tensor model working on the smallest eigenvalue of the diffusivity function for the tensor matrix of any higher order; and Ghosh et al. (2009, 2014) introduce the strategy of ternary quartic (TQ) based on Hilbert's Theorem Hilbert (1888) to parameterize the non-negative 3D kurtosis tensor by means

of a sum of three squares of quadratic forms. These ideas allow us to reconstruct the DKI which preserves the positive semi-definiteness for $W(\mathbf{g})$. However, for the upper bound, we need to evaluate the constraint at every acquisition in the computation. In DKI, parameter estimation becomes a computationally expensive constrained optimization problem. The most popular and fast methods include constrained least squares method by Tabesh et al. (2011) and the constrained weighted least squares (CWLS), see Ghosh et al. (2014); Liu (2015). For all these methods it is assumed that the observation noise is Gaussian, which however is far beyond the truth in the regime of low signal to noise ratio (SNR). In order to access information in the low SNR region, maximum likelihood estimation (MLE) based methods are also introduced in Veraart et al. (2011); Ghosh et al. (2014), where the Rician structure of the noise is accounted for, improving the accuracy of the diffusion and kurtosis tensor estimators. Liu (2015) also proposes the use of the expectation-maximization (EM) algorithm in the DKI estimation and compared the method with MLE. All these methods accounting for the Rician likelihood can only produce point estimates with frequentist confidence intervals. Furthermore, all these works assume independence between the voxels and, as a consequence, the mutually neighbouring information is completely ignored in the estimation.

In this paper, we propose a Bayesian estimation method under Rician likelihood, which considers all the three constraints in the DKI estimation. We apply the variational Bayes method (VB) to approximate the posterior probability distributions of all the parameters of interest and assess the uncertainty in the estimators. A smoothing scheme is introduced in order to access the mutual voxel-wise neighbouring information. The three main contributions of this paper thereby include: 1) Bayesian modeling in DKI; 2) A VB method for posterior approximation and for global optimization; 3) Image regularization by the Gaussian Markov Random Field (GMRF). In addition, we present a case-control study with subjects affected by Lévy Body Dementia Disease (LBD).

Section 2 overviews the DKI model and the Rician noise structure of the signal. In Section 3, we introduce the Bayesian modeling. Our VB method and imaging regularization technique are outlined in Section 4 and 5, respectively. We applied our

methods on both synthetic and real data in Section 6. The paper ends with a discussion in Section 6.

2. Theory

2.1. Preliminaries

The convenient definition of *kurtosis* of a distribution is

$$\mathrm{Kur} = \frac{M_4}{M_2^2} - 3, \tag{3}$$

where M_n denotes the n-th central moment. Diffusion weighted imaging (DWI) is based on the Fourier relationship between the signal decay and the distribution of water molecules displacement. The signal decay thus can measure the average of diffusion displacement of water molecules and is defined as the expectation of the signal phase w.r.t. the diffusion distribution function (CDF) $\mathbb{P}(\mathbf{x},t)$ of the diffusion displacement $\mathbf{x}(t) \in \mathbb{R}^3$ over the time t between the dephasing and rephasing of the gradient pulses (see Jensen et al. (2005); Mori (2007); Zhu et al. (2007); Descoteaux et al. (2011)). It can be written as

$$S/S_0 = \langle \phi(\mathbf{x}(t)) \rangle = \int_{\mathbf{x}\in\mathbb{R}^3} \phi(\mathbf{x}) d\mathbb{P}(\mathbf{x},t), \tag{4}$$

where the signal phase is defined as $\phi(\mathbf{x}) = \exp(i|\mathbf{G}|\gamma\delta\ \mathbf{g}\cdot\mathbf{x})$, with gradient scheme $(\delta,\mathbf{G})$ including the gradient amplitude $|\mathbf{G}|$, and direction $\mathbf{g} = \mathbf{G}/|\mathbf{G}|$, the duration of the gradient pulse δ, the proton gyromagnetic ratio γ and the signal intensity without diffusion weighting S_0. Here $(\mathbf{g}\cdot\mathbf{x})$ is the inner product, and in addition, the angle bracket stands for the expectation.

The characteristic function Eq. (4), can be interpreted as the characteristic function of the random vector $\mathbf{x}(t)$. We expand Eq. 4 into the Taylor series, obtaining the signal decay

$$S/S_0 = E\big(\exp(i|\mathbf{G}|\gamma\delta\ \mathbf{x}(t)\cdot\mathbf{g})\big) = \sum_{n\geq 0} \frac{(-1)^n \gamma^{2n}\delta^{2n}|\mathbf{G}|^{2n}}{(2n)!} \langle (\mathbf{x}(t)\cdot\mathbf{g})^{2n} \rangle \ \in [0,1]\,, \tag{5}$$

where the odd moments $\langle (\mathbf{x}(t)\cdot\mathbf{g})^{2n+1} \rangle$ vanish since $\mathbb{P}(\mathbf{x},t)$ is reflection symmetric. Let $\sigma^2(t)$ denote the variance of $(\mathbf{x}(t)\cdot\mathbf{g})$ which determines the width of the diffusion

distribution as the mean square travel distance of water molecules in direction $\mathbf{g}$. By Einstein's equation, $\sigma^2(t) = 2tD_{app}$, where

$$D_{app} := \mathbf{g}^T \Theta \mathbf{g} = \sum_{\ell_1,\ell_2=1}^{3} g_{\ell_1} g_{\ell_2} \Theta_{\ell_1,\ell_2},$$

is the *apparent diffusion coefficient*, see for instance Jensen et al. (2005). We expand the logarithm of Eq. (5) retaining only the terms up to 4-th order, obtaining the approximation

$$\begin{aligned}\log(S/S_0) &\simeq -\frac{\gamma^2\delta^2|\mathbf{G}|^2}{2}\langle(\mathbf{x}(t)\cdot\mathbf{g})^2\rangle + \frac{\gamma^4\delta^4|\mathbf{G}|^4}{4!}\big(\langle(\mathbf{x}(t)\cdot\mathbf{g})^4\rangle - 3\langle(\mathbf{x}(t)\cdot\mathbf{g})^2\rangle^2\big)\\ &= -\gamma^2\delta^2|\mathbf{G}|^2 tD_{app} + \frac{\gamma^4\delta^4|\mathbf{G}|^4 t^2 D_{app}^2}{6}\left(\frac{\langle(\mathbf{x}(t)\cdot\mathbf{g})^4\rangle}{4t^2D_{app}^2} - 3\right).\end{aligned}$$

This justifies the DKI signal model

$$S(b)/S_0 = \exp\left(-bD_{app} + \frac{1}{6}b^2 D_{app}^2 K_{app}\right) \quad \in \mathbb{R}, \tag{6}$$

see Jensen et al. (2005); Helpern et al. (2009); Jensen and Helpern (2010), derived by assuming that the pulse during time δ is short enough, so that the Stejskal-Tanner sequence has the approximation $b \approx t(\gamma\delta|\mathbf{G}|)^2$, see for instance Qi et al. (2009); Tabesh et al. (2011); Steven et al. (2014). Following Eq. (6),

$$K_{app} = \frac{\langle(\mathbf{x}(\mathbf{t})\cdot\mathbf{g})^4\rangle}{4t^2D_{app}^2} - 3 = \left(\frac{\overline{\mathrm{tr}(\Theta)}}{D_{app}}\right)^2 W_{app}, \tag{7}$$

is the *apparent diffusion kurtosis*, $\overline{tr(\Theta)} = tr(\Theta)/3$ is the *mean diffusivity*, and

$$W_{app} = \sum_{\ell_1,\ell_2\ell_3\ell_4=1}^{3} W_{\ell_1,\ell_2\ell_3\ell_4} g_{\ell_1} g_{\ell_2} g_{\ell_3} g_{\ell_4}$$

is the totally symmetric 4th-order *kurtosis tensor* W, see e.g., Qi et al. (2009); Tabesh et al. (2011); Steven et al. (2014). Replacing K_{app} in Eq. (6) by W_{app} in Eq. (7), we get the DKI model in Eq. (1) in coincidence with the one described in Section 1. Hence the 2nd (# 2) and 3rd (# 3) constraints in DKI corresponding to K_{app} is

$$0 \le K_{app} \le 3/(bD_{app}),$$

which is equivalent to the constraints in Eq. (2).

2.2. Rician Likelihood and Data Augmentation

We first introduce the noise model for the raw MRI-acquisitions. The noise in MRI is complex-valued and is composed of two *i.i.d*. Gaussian random variables, ε_r and ε_i, with zero mean and variance σ^2 specified for the real and imaginary components, respectively. The joint density of the MRI-noise is expressed by

$$p_{S,\sigma^2}(\varepsilon_r,\ \varepsilon_i) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\varepsilon_r^2+\varepsilon_i^2}{2\sigma^2}\right), \tag{8}$$

see Andersen (1996); Koay and Basser (2006); Zhu et al. (2007). The MRI-measurement Y is the magnitude of the signal intensity $S \geq 0$ corrupted by complex-valued noise, expressed as

$$Y = |S+\varepsilon| = \sqrt{(S+\varepsilon_r)^2+\varepsilon_i^2}.$$

Let φ be the phase data defined as $\varphi := \arg\left(S+\varepsilon_r+i\varepsilon_i\right) \in (-\pi,\pi]$ such that $S+\varepsilon_r = Y\cos\varphi$ and $\varepsilon_i = Y\sin\varphi$. By the change of variables formula the joint density of Y and φ with parameters S and σ^2 is given by

$$\begin{aligned} p_{S,\sigma^2}(y,\varphi) &= \frac{y}{2\pi\sigma^2}\exp\left(-\frac{(y\cos\varphi - S)^2 + y^2(\sin\varphi)^2}{2\sigma^2}\right) \\ &= \frac{y}{2\pi\sigma^2}\exp\left(-\frac{y^2+S^2-2yS\cos\varphi}{2\sigma^2}\right) \\ &= p_{S,\sigma^2}(y)p_{S,\sigma^2}(\varphi|y). \end{aligned} \tag{9}$$

The marginal density of Y has a Rician distribution with likelihood function

$$p_{S,\sigma^2}(y) = \frac{y}{\sigma^2}\exp\left(-\frac{y^2+S^2}{2\sigma^2}\right)I_0\left(\frac{yS}{\sigma^2}\right)\mathbb{1}(y\geq 0), \tag{10}$$

see Henkelman (1985); Gudbjartsson and Patz (1995). $I_0(\cdot)$ is the zero-order modified Bessel function of the first kind and $\mathbb{1}(\cdot)$ is the indicator function. The conditional density

$$p_{S,\sigma^2}(\varphi|y) = \frac{1}{2\pi I_0(Sy/\sigma^2)}\exp\left(\frac{yS}{\sigma^2}\cos\varphi\right), \quad \varphi\in(-\pi,\pi], \tag{11}$$

is an instance of the Von Mises distribution on the unit circle symmetric around zero, see Fisher (1993). Note that, although in theory the zero magnitude is obtained with zero probability, in reality we can still acquire $y = 0$ due to numerical truncation. In such a case, the noise terms only contain the real Gaussian component and the data have a Gaussian likelihood.

2.3. Parametrization in DKI

We parametrize the DKI model as follows: for an acquisition with given b-value and gradient $\mathbf{g}$, the log-signal is given by

$$\log S = \log S_0 - b\mathbf{g}^\top \Theta \mathbf{g} + \sum_{i=1}^{3} \left(b\mathbf{g}^\top \Psi^{(i)} \mathbf{g}\right)^2, \tag{12}$$

where S_0 is the baseline unweighted signal, Θ and $\Psi^{(i)}, i = 1,2,3$, are 3×3 symmetric matrices, parametrizing the kurtosis tensor as a ternary quartic as in Ghosh et al. (2009). Equivalently

$$\log S - \log S_0 = Z\theta + Z\psi\psi^\top Z^\top = Z\theta + \mathrm{Trace}\left(\psi^\top Z^\top Z\psi\right) = Z\theta + \parallel Z\psi \parallel^2, \tag{13}$$

where

$$Z = Z(b,\mathbf{g}) = -b\left(g_1^2, g_2^2, g_3^2, 2g_1g_2, 2g_1g_3, 2g_2g_3\right), \tag{14}$$

$$\theta = \mathrm{vec}(\Theta) = (\Theta_{11}, \Theta_{22}, \Theta_{33}, \Theta_{12}, \Theta_{13}, \Theta_{23})^\top$$

and ψ is a 6×3 symmetric matrix with colums

$$\psi_{\bullet i} = \mathrm{vec}(\Psi^{(i)}) = (\Psi^{(i)}_{11}, \Psi^{(i)}_{22}, \Psi^{(i)}_{33}, \Psi^{(i)}_{12}, \Psi^{(i)}_{13}, \Psi^{(i)}_{23})^\top,\ i = 1,2,3.$$

When the diffusion tensor Θ is positive definite, it can be further parametrized by using its Cholesky decomposition $\Theta = LL^T$, with L upper triangular. Correspondingly,

$$\theta = \theta(L) = (L_{11}^2, L_{12}^2 + L_{22}^2, L_{13}^2 + L_{23}^2 + L_{33}^2, L_{11}L_{12}, L_{11}L_{13}, L_{12}L_{13} + L_{22}L_{23}). \tag{15}$$

Combining Eq. (9) and Eq. (13), we then get the augmented log-likelihood for the DKI parameters under the complex noise model:

$$\begin{aligned} &\log p_{\theta,\psi,S_0,\sigma^2}(Y,\varphi) = \log(Y) - \log(2\pi) - \log(\sigma^2) \\ &- \frac{1}{2\sigma^2}\left\{Y^2 + S_0^2 \exp\left(2Z\theta + 2 \parallel Z\psi \parallel^2\right) - 2\exp\left(Z\theta + \parallel Z\psi \parallel^2\right) Y S_0 \cos\varphi\right\}, \end{aligned} \tag{16}$$

where the exponential function is defined componentwise.

3. Bayesian modeling

In Bayesian theory, all the unknown parameters are treated as random variables with an assumed prior distribution. This is the essential difference between the Bayesian and the frequentist methods. Statistical inference is then based on the posterior distribution conditionally on the observed data, derived by the Bayes rule

$$p(\xi|y) = \frac{p(\xi,y)}{p(y)} = \frac{p(\xi)p(y|\xi)}{p(y)}, \tag{17}$$

see for instance Lindley (1972); Gelman et al. (2014); Berger (2013), where the posterior density $p(\xi|y)$ depends not only on the likelihood but also on the prior, encoding the pre-existent knowledge about the unknown parameters.

The variational Bayes method approximates the posterior without the computation of the normalizing constant $p(y)$. To apply this method in DKI, we need to construct a Bayesian model by specifying the prior for the parameters. For the single voxel case, the parameters are θ, ψ, σ^2 and S_0. A Bayesian hierarchical prior model is constructed as follows:

- S_0 has a constrained Gaussian prior

$$\pi(S_0) = \frac{1}{\Phi(\mu_0/\eta_0)} \exp\left(-\frac{1}{2\eta_0^2}(S_0-\mu_0)^2\right) \mathbb{1}(S_0 \geq 0).$$

- The noise parameter σ^2 has an inverse Gamma prior with density

$$\pi(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-2(\alpha+1)} \exp\left(-\beta/\sigma^2\right)$$

  with $\alpha, \beta > 0$. We can also use the scale invariant improper prior $\pi(\sigma^2) \propto \sigma^{-2}$, corresponding to $\alpha = \beta = 0$. Although the prior does not integrate on $[0,\infty)$, the posterior of σ^{-2} will be integrable when θ has a proper prior and the number of acquisitions is $m >$ rank(Z), see Gasbarra et al. (2014) for the details.

- We specify independent zero-mean rotation invariant Gaussian priors for the 3×3 symmetric matrices Θ and $\Psi^{(i)}$, $i = 1,2,3$. The prior for Θ is

$$\pi(\Theta) = \frac{\eta^{5/2}\sqrt{\eta+3\lambda}}{(\pi\sqrt{2})^3} \exp\left(-\frac{1}{2}\left(\eta \text{Trace}(\Theta^2) + \lambda\left\{\text{Trace}(\Theta)\right\}^2\right)\right) \tag{18}$$

as in Basser and Pajevic (2003), where $\eta > 0$, $\lambda > -2\eta/3$ are hyperparameters. Equivalently, $\theta = \mathrm{vec}(\theta)$ and the column vectors $\psi_{\bullet i} = \mathrm{vec}(\Psi^{(i)}), i = 1,2,3$, have *i.i.d.* zero-mean Gaussian priors with precision matrix

$$\Omega = \begin{pmatrix} \lambda+2\eta & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+2\eta & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+2\eta & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\eta & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\eta & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\eta \end{pmatrix}. \tag{19}$$

The Bayes formula combines the above log priors with the augmented log-likelihood Eq. (16) into the log-posterior and gives

$$\begin{aligned} &\log p(\theta,\psi,S_0,\sigma^2,\varphi|Y) = \text{const.} + \log\pi(S_0) - \frac{1}{2}\theta^\top\Omega\theta - \frac{1}{2}\mathrm{Trace}\big(\psi^\top\Omega\psi\big) \\ &-(m+1)\log(\sigma^2) - \frac{S_0^2}{2\sigma^2}\sum_{j=1}^{m}\big\{Y_j^2 + \exp\big(2Z_j\theta + 2\parallel Z_j\psi\parallel^2\big)\big\} \\ &+\frac{S_0}{\sigma^2}\sum_{j=1}^{m}\exp\big(Z_j\theta + \parallel Z_j\psi\parallel^2\big)Y_j\cos\varphi_j, \end{aligned} \tag{20}$$

where m is the number of acquisitions and $Z_j = Z(b_j, \mathbf{g_j})$, $j = 1,\ldots,m$.

4. Variational Bayes approximation

The idea of the mean-field variational Bayes framework (VB) is to approximate the posterior distribution of the parameter $\xi = (\xi_1,\ldots,\xi_n)$ given the response y,

$$p(\xi|y) = p(\xi)p(y|\xi)/p(y),$$

simply by a product of probability distributions

$$\widehat{q}(\xi) = \widehat{q}_1(\xi_1)\widehat{q}_2(\xi_2)\cdots\widehat{q}_n(\xi_n),$$

corresponding to independence of the components ξ_k under the approximative distribution $\widehat{q}(\xi)$. The VB approximation is found by minimizing the Kullback-Leibler (KL) divergence

$$\mathrm{KL}\big(p(\cdot|y)\parallel q(\cdot)\big) = \int q(\xi)\log\left(\frac{q(\xi)}{p(\xi|y)}\right)d\xi, \tag{21}$$

using recursion, see Kullback and Leibler (1951) for instance. The VB-marginal recursions are

$$\widehat{q}_k^{(t+1)}(\xi_k) \propto \int \log p(y,\xi) \prod_{h\neq k} \widehat{q}_h^{(t)}(\xi_h) d\xi_1 \dots d\xi_{k-1}\, d\xi_{k+1} \dots d\xi_n, \tag{22}$$

as in Šmídl and Quinn (2006). The detailed explanation and derivation of the VB updating algorithm can be found in Jaakkola and Jordan (2000); Ormerod and Wand (2010) .

4.1. VB-marginals in constrained DKI

We shall use the VB method to approximate the joint posterior distribution

$$p(\sigma^2, \theta, \psi, S_0, \varphi_1, \dots, \varphi_m | Y_1, \dots, Y_m), \tag{23}$$

whose logarithm is given in Eq. (20), by a product distribution with factorization

$$\widehat{q}_\theta(\theta)\widehat{q}_\psi(\psi)\widehat{q}_{S_0}(S_0)\widehat{q}_{\sigma^2}(\sigma^2)\widehat{q}_1(\varphi_1)\dots\widehat{q}_m(\varphi_m).$$

The VB-marginals of the unknown parameters in DKI are calculated by Eq. (22). In what follows the VB-expectations, denoted by the angle bracket $\langle\cdot\rangle$, can be computed as shown at the end of this subsection.

We start from the VB-marginal of the baseline signal intensity given by

$$\begin{aligned}
&\widehat{q}(S_0) \propto \\
&\mathbb{1}(S_0 \geq 0) \exp\Bigg(-\frac{(S_0-\mu_0)^2}{2\eta_0^2} - \frac{S_0^2\langle\sigma^{-2}\rangle}{2} \sum_{j=1}^m \langle \exp(2Z_j\theta)\rangle \langle \exp(2 \parallel Z_j\psi \parallel^2)\rangle \\
&+ S_0\langle\sigma^{-2}\rangle \sum_{j=1}^m Y_j \langle\cos\varphi_j\rangle \langle \exp(Z_j\theta)\rangle \langle \exp(\parallel Z_j\psi \parallel^2)\rangle \Bigg),
\end{aligned}$$

which is the constrained Gaussian density with mean

$$\widehat{\mu} = \left(\mu_0\eta_0^{-2} + \langle\sigma^{-2}\rangle \sum_{j=1}^m Y_j \langle\cos\varphi_j\rangle \langle \exp(Z_j\theta)\rangle \langle \exp(\parallel Z_j\psi \parallel^2)\rangle \right) \widehat{\eta}^2$$

and variance

$$\widehat{\eta}^2 = \left(\eta_0^{-2} + \langle\sigma^{-2}\rangle \sum_{j=1}^m \langle \exp(2Z_j\theta)\rangle \langle \exp(2 \parallel Z_j\psi \parallel^2)\rangle \right)^{-1}$$

constrained on $\mathbb{R}^+$. In order to derive the other VB-marginals, we compute the first two moments of the constrained Gaussian variable $S_0 \sim \mathcal{N}\left(\widehat{\mu}, \widehat{\eta}^2\right)$ given by

$$\left\langle S_0 | \mathbb{1}(S_0 \geq 0) \right\rangle = \widehat{\mu} + \widehat{\eta} \frac{\phi(\widehat{\mu}/\widehat{\eta})}{\Phi(\widehat{\mu}/\widehat{\eta})}$$

and

$$\left\langle S_0^2 | \mathbb{1}(S_0 \geq 0) \right\rangle = \widehat{\mu}^2 + \widehat{\eta}^2 + \widehat{\mu}\widehat{\eta} \frac{\phi(\widehat{\mu}/\widehat{\eta})}{\Phi(\widehat{\mu}/\widehat{\eta})},$$

where $\phi(t)$ and $\Phi(t)$ are the density (PDF) and cumulative distribution (CDF) functions of the standard Gaussian distribution, respectively. The expectations above can be derived either by using the Fubini theorem, see the details in Appendix C, or by using the moment generating function of the one-dimensional constrained Gaussian distribution, see Nadarajah and Kotz (2008).

The VB-marginal of φ_j is given by

$$\widehat{q}(\varphi_j) \propto \exp\Big(Y_j \cos\varphi_j \langle \sigma^{-2} \rangle \langle S_0 \rangle \langle \exp\left(Z_j \theta\right) \rangle \langle \exp\left(\| \, Z_j \psi \, \|^2 \right) \rangle \Big), \ \varphi_j \in (-\pi, \pi],$$

which is a symmetric Von Mises distribution with parameter

$$\widehat{\kappa}_j = Y_j \langle \sigma^{-2} \rangle \langle S_0 \rangle \langle \exp\left(Z_j \theta\right) \rangle \langle \exp\left(\| \, Z_j \psi \, \|^2 \right) \rangle, \quad \text{such that } \left\langle \cos\varphi_j \right\rangle = \frac{I_1\left(\widehat{\kappa}_j\right)}{I_0\left(\widehat{\kappa}_j\right)},$$

where I_z is the modified Bessel function of the first kind.

The VB marginal of σ^2 is an inverse gamma distribution with shape parameter m and rate parameter

$$\begin{aligned} \hat{\nu} = & \frac{1}{2} \sum_{j=1}^{m} \Big\{ Y_j^2 + \langle S_0^2 \rangle \langle \exp\left(2 Z_j \theta\right) \rangle \langle \exp\left(2 \, \| \, Z_j \psi \, \|^2\right) \rangle \\ & - 2 Y_j \langle \cos\varphi_j \rangle \langle S_0 \rangle \langle \exp\left(Z_j \theta\right) \rangle \langle \exp\left(\| \, Z_j \psi \, \|^2 \right) \rangle \Big\} \end{aligned}$$

with density

$$\widehat{q}(\sigma^2) \propto \sigma^{-2(m+1)} \exp\left(-\widehat{\nu} \sigma^{-2}\right),$$

such that $\langle \sigma^{-2} \rangle = m/\hat{\nu}$.

4.2. Laplace approximation and the delta method

The exact VB-marginals of the diffusion tensor parameters up to a normalizing constant is given by

$$\widehat{q}(\theta) \propto \exp\Big(-\frac{1}{2}\theta^{\top}\Omega\theta - \frac{\langle S_0^2\rangle\langle\sigma^{-2}\rangle}{2}\sum_{j=1}^{m}\exp\big(2Z_j\theta\big)\big\langle\exp\big(2\parallel Z_j\psi\parallel^2\big)\big\rangle + \langle S_0\rangle\langle\sigma^{-2}\rangle\sum_{j=1}^{m}\exp\big(Z_j\theta\big)Y_j\langle\cos\varphi_j\rangle\big\langle\exp\big(\parallel Z_j\psi\parallel^2\big)\big\rangle\Big), \tag{24}$$

and for the kurtosis tensor parameters we have

$$\widehat{q}(\psi) \propto \exp\Big(-\frac{1}{2}\mathrm{Trace}(\psi^{\top}\Omega\psi) - \frac{\langle S_0^2\rangle\langle\sigma^{-2}\rangle}{2}\sum_{j=1}^{m}\exp\big(2Z_j\psi\psi^{\top}Z_j^{\top}\big)\big\langle\exp\big(2Z_j\theta\big)\big\rangle + \langle S_0\rangle\langle\sigma^{-2}\rangle\sum_{j=1}^{m}\exp\big(Z_j\psi\psi^{\top}Z_j^{\top}\big)Y_j\langle\cos\varphi_j\rangle\big\langle\exp\big(Z_j\theta\big)\big\rangle\Big). \tag{25}$$

These exact VB-marginals are non-standard distributions which can not be integrated analytically. Direct implementation of the VB-model is not possible, and a further numerical approximation is needed.

We impose the following restriction on the VB-marginals of the tensor and kurtosis parameters θ and ψ by assuming that the marginals of θ and ψ have an approximation in terms of multivariate Gaussian distributions with mean $\widehat{\theta}$ and $\widehat{\psi}$, respectively, satisfying the DKI model constraints:

a) $\widehat{\theta}$ corresponds to a positive definite symmetric matrix $\widehat{\Theta}$ with the Cholesky decomposition $\hat{\Theta} = L^{\top}L$.

b) $Z_j\widehat{\theta} + 2\parallel Z_j\widehat{\psi}\parallel^2 \leq 0 \quad \forall j = 1,\ldots,m.$

The mean tensor and kurtosis parameters $\hat{\theta}$ and $\hat{\psi}$ of the approximative Gaussian marginals are found by maximizing the exact VB-marginals (24) and (25) respectively under the DKI constraints, keeping the other VB-marginals fixed, and the covariances are determined by the second order terms in the Laplace approximation.

Since the mapping $L \to \theta(L)$ in Eq. (15) between upper-triangular matrices L and the positive tensor parameters θ is one-to-one, by the delta method (see e.g., Casella and Berger (2002)) the image probability distribution of L has also a Gaussian approximation.

The Laplace approximation of $q(L)$ is approximated by the Gaussian distribution $\mathcal{N}(\widehat{L},\widehat{\Sigma}_L)$, where $\widehat{L}$ is the $q(L)$ mode and $\widehat{\Sigma}_L$ is the Hessian matrix at the mode. By applying the delta-method we obtain the Laplace approximation of $q(\theta_D)$, given by $q(\theta_D) \xrightarrow{d} \mathcal{N}(\widehat{L}^T\widehat{L},\ \widehat{\Sigma}_D)$ with $\widehat{\Sigma}_D = \widehat{J}_L^T\widehat{\Sigma}_L\widehat{J}_L$ and

$$J_L = \left(\frac{\partial\theta_i}{\partial L_j}\right)_{i,j=1,\cdots,6} = \begin{pmatrix} 2L_{11} & & & & & \\ & 2L_{22} & & 2L_{12} & & \\ & & 2L_{33} & & 2L_{13} & 2L_{23} \\ L_{12} & & & L_{11} & & \\ L_{13} & & & & L_{11} & \\ & L_{23} & & L_{13} & L_{12} & L_{22} \end{pmatrix}. \tag{26}$$

In the computation of the other VB-marginals, we need the exponential moments

$$\left\langle \exp(tZ_j\theta)\right\rangle = \exp\left(tZ_j\widehat{\theta} + \frac{t^2}{2}Z_j\widehat{\Sigma}_\theta Z_j^\top\right), \quad j = 1,\ldots,m,\ t = 1,2\,.$$

We use the same idea to construct the approximative Gaussian VB-marginal for the kurtosis parameters, with $\mathrm{vec}(\psi) \sim \mathcal{N}\left(\mathrm{vec}(\widehat{\psi}),\widehat{\Sigma}_\psi\right)$. In the expression for the other VB-marginals for $t = 1,2$ we need the exponential moments

$$\begin{aligned}&\left\langle \exp\left(t \parallel Z_j\psi \parallel^2\right)\right\rangle \\ &= \det(\mathrm{Id} - 2t\widehat{\Sigma}_\psi A_j)^{-\frac{1}{2}} \exp\left(\frac{1}{2}\mathrm{vec}(\widehat{\psi})^\top\widehat{\Sigma}_\psi^{-1}\left\{(\mathrm{Id} - 2t\widehat{\Sigma}_\psi A_j)^{-1} - \mathrm{Id}\right\}\mathrm{vec}(\widehat{\psi})\right),\end{aligned}$$

where

$$A_j = \mathrm{Id}_{3\times3} \otimes \left(Z_j^\top Z_j\right) \in \mathbb{R}^{18\times18}$$

is a tensor product of matrices and Id is the identity matrix, and we assume that the matrices

$$\left(\mathrm{Id} - 4\widehat{\Sigma}_\psi A_j\right) \in \mathbb{R}^{18\times18} \quad j = 1,\ldots,m,$$

are positive definite, see Appendix B for a detailed explanation.

In practice this condition can be violated when the data do not contain enough information and the variances in the marginal distribution of the kurtosis parameters are too large. In such a case it may be necessary to collect more data.

The stopping criteria. The KL divergence

$$0 \le \mathrm{KL}(\widehat{q}(\xi)||p(\xi|y)) = \int \log\left(\frac{\widehat{q}^{(t)}(\xi)}{p(\xi|y)}\right)\widehat{q}^{(t)}(\xi)d\xi =$$
$$\sum_{i=1}^{n}\int \log(\widehat{q}_i(\xi_i))\widehat{q}_i(\xi_i)d\xi_i - \int \log(p(\xi_1,\ldots,\xi_n,y))\prod_{i=1}^{n}\widehat{q}_i(\xi_i)d\xi_i + \log p(y)$$

does not increase between consecutive VB-updates. Therefore, the VB algorithm can be stopped when the decrease of the KL divergence is negligible, see Ormerod and Wand (2010); Šmídl and Quinn (2006). For our model, up to an additive constant, the KL divergence is given by

$$\mathrm{KL}(\widehat{q}(\xi)||p(\xi|y)) = \mathrm{const} + \frac{\widehat{\mu}}{2\widehat{\eta}}\frac{\phi(\widehat{\mu}/\widehat{\eta})}{\Phi(\widehat{\mu}/\widehat{\eta})} - \log\big(\widehat{\eta}\Phi(\widehat{\mu}/\widehat{\eta})\big) - \frac{1}{2}\log|\Sigma_\theta|$$
$$-\frac{1}{2}\log|\Sigma_\psi| + \frac{1}{2}\mu_\theta^\top\Omega\mu_\theta + \frac{1}{2}\mathrm{Trace}(\Omega\Sigma_\theta) + \frac{1}{2}\mathrm{Trace}\big(\mu_\psi^\top\Omega\mu_\psi\big) + \frac{1}{2}\mathrm{Trace}(\Sigma_\psi:\Omega)$$
$$+\frac{\langle S_0^2\rangle\langle\sigma^{-2}\rangle}{2}\sum_{j=1}^{m}\big\{Y_j^2 + \langle\exp\big(2Z_j\theta\big)\rangle\langle\exp\big(2\parallel Z_j\psi\parallel^2\big)\rangle\big\},$$

see the detailed calculation in Appendix D.

4.3. Upper bound of K_{app} and nonlinear optimization

In order to find numerically the VB marginals of the diffusion and kurtosis tensor parameters θ and ψ we need to find the modes of their Laplace approximations. We also need to take into account the upper bound on K_{app}, which yields nonlinear constraints

$$2\,Z_j\psi\psi^\top Z_j^\top + Z\theta \le 0 \qquad j = 1,\ldots,m$$

at every acquisition.

Constrained optimization w.r.t. θ,ψ is achieved by using the method proposed in Liu (2015). Note that in this approach only the modes $\widehat{\theta},\widehat{\psi}$ of the Gaussian VB-marginals satisfy the constraints, and the VB-expectations are computed by integrating from these unconstrained distributions.

5. Image regularization

The brain white matter has highly organized structure and diffusion tensor orientations coincide with axon directions. Water molecules diffuse mostly along the direction

of the underlying axon fibres and, as a consequence, the tensors at different locations should not be considered as statistically independent. The correlation between two tensors / blocks of tensors depends on the distance where they locate physically. In order to estimate consistent diffusion images, it is useful to use at every single voxel the information from neighbouring voxels, especially when data are corrupted and / or contain missing observations. The parameters for multiple voxels are then $\theta(v), \psi(v), \sigma_v^2$ and $S_0(v), v \in V$, where $V \subseteq \mathbb{Z}^3$ is the region of interest (ROI). To simplify the notations, we omit the subscripts v denoted the position of a voxel in σ_v^2 and $S_0(v)$. We thus consider the neighbourhood relation $v \sim w$ on the values $(\theta(v) : v \in \partial W)$ and specify a prior contribution, a joint density of $\theta(v)$ and $\theta(w)$ to replace Eq. (18), which is given by

$$\begin{aligned}&\log \pi(\theta(w) : w \in W; \theta(v), v \in \partial W) \\ &= \text{const} \; - \frac{\rho}{2} \sum_{v \sim w : v \in W, w \in \overline{W}} (\theta(v) - \theta(w))^\top \Omega (\theta(v) - \theta(w)) - \frac{1}{2} \sum_{v \in W} \theta(v)^\top \Omega \theta(v),\end{aligned} \tag{27}$$

where $\rho \geq 0$ tunes the strength of dependence between the neighbour tensors.We denote the exterior boundary of W by

$$\partial W := \{w \in V \setminus W : \; \exists v \in W \text{ with } w \sim v\}$$

and set $\overline{W} := W \cup \partial W$, see Gasbarra et al. (2014); Kaipio and Somersalo (2006).
We extend the idea and construct the GMRF prior for the 4th-order positive tensor using the kurtosis $\psi(v)$ as an illustration, which is given by

$$\begin{aligned}&\log \pi(\text{vec}(\psi(w)) : w \in W; \theta(v), v \in \partial W) = \text{const} \; - \frac{\rho}{2} \sum_{v \sim w : v \in W, w \in \overline{W}} \\ &\text{Trace}\Big((\psi(v) - \psi(w))^\top \Omega (\psi(v) - \psi(w)) \Big) - \frac{1}{2} \sum_{v \in W} \text{Trace}\Big(\psi(v)^\top \Omega \psi(v) \Big).\end{aligned} \tag{28}$$

In this way regularization is implementable simultaneously in the VB estimation.

The regularization parameters η and λ are assumed to be known. Alternatively, we could treat them as unknown parameters with a given prior, and extend the VB algorithm computing their VB-marginals.

6. Results

The result section contains two parts. In the first part, we apply the proposed method in the study of simulated data for which we know the ground truth (GT). The 2nd part is a case-control study on real human brain data.

6.1. Simulated data

We use simulated data publicly available at
`http://projects.iq.harvard.edu/sparcdmri/Challenge_Data`.
This data are acquired based on a single slice of a physical phantom with dimension of 13×16 and with thickness of 7 mm. TE/TR (Echo Time /Repetition Time) are 41/3400 ms. The b-values are 1000, 2000, 3000 and 60 gradient directions per shell are used to acquire the diffusion-weighted signal. The encoding directions distribute unequally on the three shells. The data also include one measurement with $b = 0$. The b-value increases in each of 60 directions from 1000 to $3000 s/mm^2$. The number of acquisitions for each voxel is 181. There are 208 voxels in total, where we masked out 40 voxels that have isotropic diffusion and study only the remaining 168 voxels. The ground truth of FA in a single fiber region is around 0.8, and the average signal to noise ratio (SNR) is about 9.5. More information about the data can be found in Ning et al. (2015).

We first plot the spatial distribution of constraint violations (CV) on the baseline image from the measurements with $b = 0$. Fig. 1 points out there are 5 voxels that violate the constraints #3 and 4 violate #1, in which four of them overlap. We evaluated CV in all the 60 encoding directions from the results by the unconstrained weighted least squares method (WLS) and found out that all voxels violate constraint #2 with intensity between $[0,\ 38]$. For all the 168 voxels, the percentage of violation is between $43.33\% \sim 63.33\%$.

We then applied the proposed VB method without and with regularization for the estimation in DKI, respectively. In the scheme of regulation, we chose the block size to be 1, that is, the neighbourhood of a voxel is constructed by its nearest neighbours in 3D lattice. We also fixed the hyperparameters as $\lambda = 0$ and $\eta = 1$.

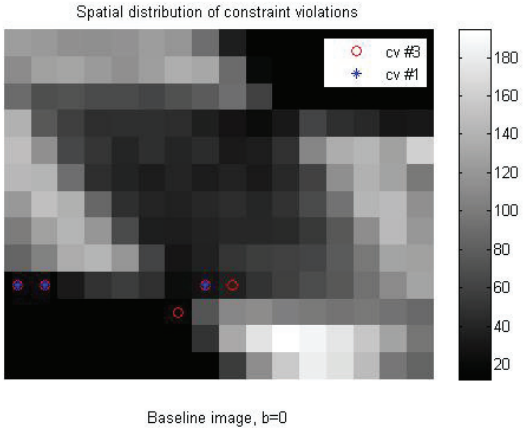


Figure 1: Spatial layout of the voxels which violate the constraint(s). The red circles indicate CV #3 and the blue stars depict CV #1. All the voxels violate constraint #2 and are not shown in this figure.

We compute the orientation diffusion functions (ODF) at each voxel under the DKI model and show ODF in Fig. 2. It reveals that the image in Fig. 2b, obtained by the regularization schemes, is much smoother than the one in Fig. 2a resulting from the independent VB scheme. In order to evaluate the accuracy of the proposed method and

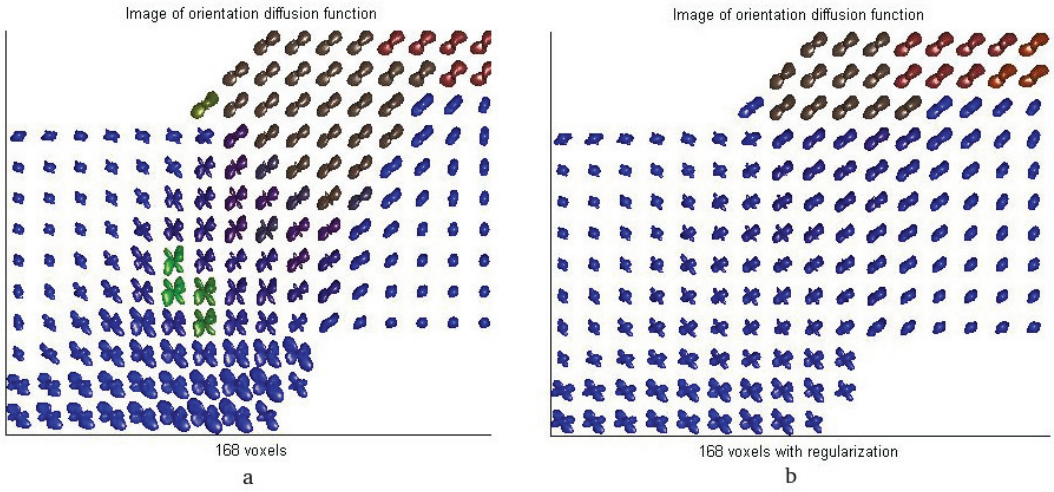


Figure 2: Image of the orientation diffusion functions (ODF) at each voxel with voxel size 168. Fig. 2a and Fig. 2b represent the results from the independent and the smooth VB schemes, respectively. The ellipsoid indicates that there is one fiber bundle at a voxel and the cross shape describes two fiber bundles at one voxel. The images are plotted with the MATLAB fanDTasia ToolBox by Barmpoutis and Vemuri (2010).

to compare the performance of the both schemes, we analyse the diffusion anisotropy of water molecules at each voxel using mean of diffusivity (MD), fractional anisotropy (FA) and the diffusion and the degree of non-Gaussianity of diffusion displacement distribution by mean kurtosis (MK). The scalar metrics are computed from the estimates of Θ and W with formula

$$\mathrm{MD} = \frac{1}{3}\mathrm{Trace}(\Theta), \qquad \mathrm{FA} = \sqrt{\frac{3\mathrm{Var}(x)}{2(x_1^2+x_2^2+x_3^2)}},$$
$$\mathrm{MK} = \frac{1}{4\pi}\int_{\mathscr{S}^2} W(\mathbf{g})d(\mathbf{g}),$$

where x_1, x_2 and x_3 are eigenvalues of tensor matrix Θ and MK is defined as the average of the observed kurtosis (and in practice, we use the estimated kurtosis) over all the directions on the unit sphere, see Tabesh et al. (2011). The calculation of MD, FA and MK were implemented in MATLAB and are visualized in Fig. 3 (MD), Fig. 4 (FA) and Fig. 5 (MK), respectively. The MK map retrieved from the regularized scheme shows higher degree of deviation from the Gaussian distribution than that obtained from the independent scheme, and vice versa in terms of the MD-parametric maps. This is in coincide with GT that the number of fiber bundles in the ROI is taken the values among $(0,\ 1,\ 2,\ 3)$, although in DKI we only consider the number of fiber bundles up to 2. The FA maps show less difference between the two schemes. The average values of FA from the regularized scheme are 0.8472 and from the independent scheme 0.8545. We list the mean values of the scalar metrics in Table 1, where we also compute the mean of SNR (S_0/σ). It is apparent that the mean of the estimated SNR values differ between the two schemes, and the value from the regularized scheme is underestimated and the scheme without regularization is overestimated, but they are both close to GT (9.0).

Table 1: Mean value of estimated scalar statistics over the ROI.

Mean values	MD ($\times mm^2/s$)	FA	MK	SNR (S_0/σ)
independence	5.3557	0.8545	0.1510	9.1258
regularization	2.7021	0.8472	0.6333	8.8512

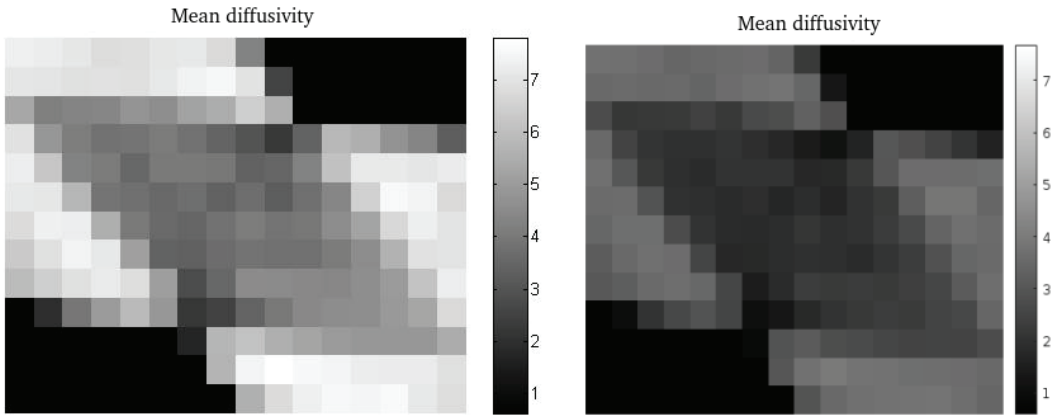


Figure 3: The MD maps with voxel size 168. Fig. 3a describes the estimates without regularization with values in [1, 7] $\times 10^{-3} mm^2/s$. Fig. 3b represent the results from the regularized VB scheme with values in [0.5, 3.5] $\times 10^{-3} mm^2/s$.

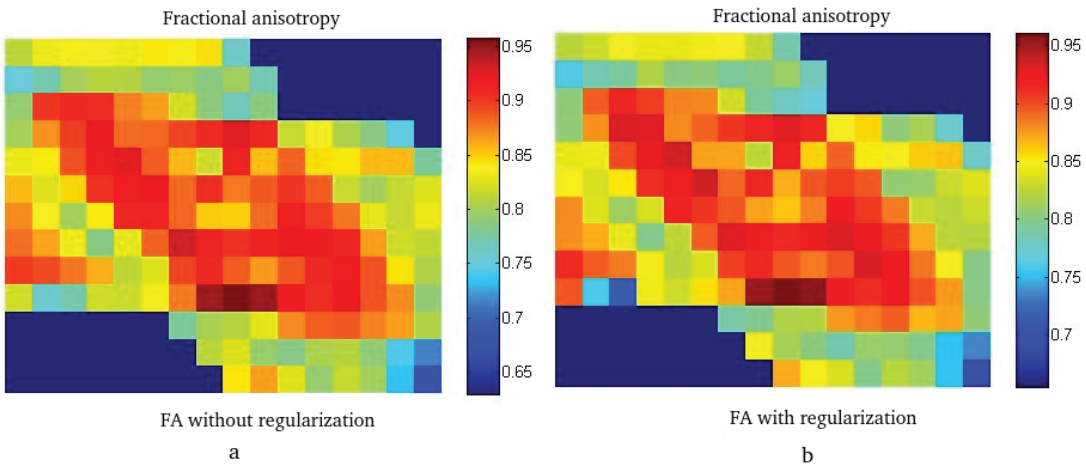


Figure 4: The FA maps with voxel size 168. Fig. 4a and Fig. 4b show the VB-estimators without and with regularization, respectively.

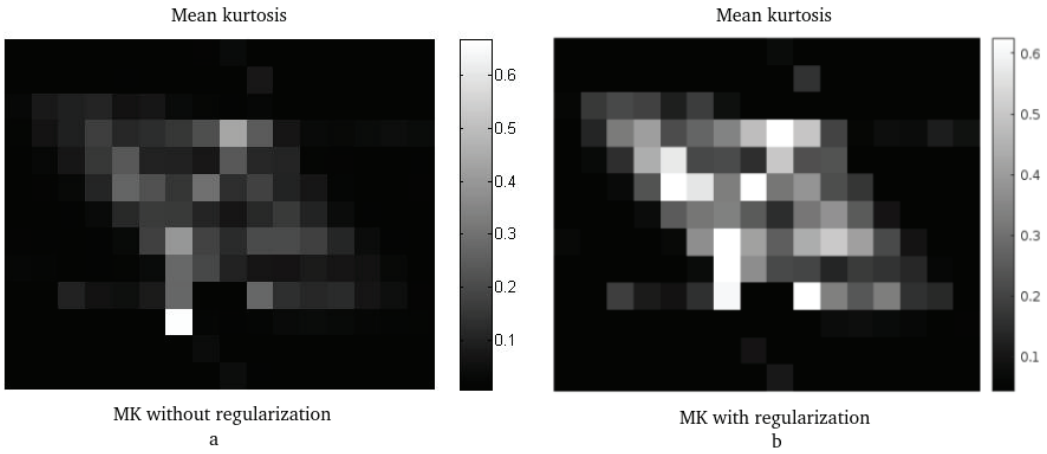


Figure 5: The MK maps with voxel size 168. Fig. 5a and Fig. 5b show the VB-estimators without and with regularization, respectively.

6.2. Real data

In the follow-up, we applied the proposed method to real data of human brain data from three subjects. The first two were healthy controls, and the third one was a case of LBD. The three datasets were acquired from multiple shells with the same gradient scheme, where the b-values were varying in the range 62, 249, 560, 996, 1556, 2240 s/mm^2 and 32 distinct gradient directions have been used, see Appendix A. All the datasets included one measurement with zero b-value. The data were collected by a Philips Achieva 3.0 Tesla MR-scanner from roughly same ROI including the corpus callosum, and the image resolution was 128×128 pixels of size 1.875×1.875 mm^2. The first dataset contains five consecutive axial slices with thickness $5mm$, and the value of TE/TR is $59.5ms/7084.4ms$. The other two datasets from the same age group contain five consecutive axial slices with thickness $4mm$, and TE/TR is $100ms/25083ms$.

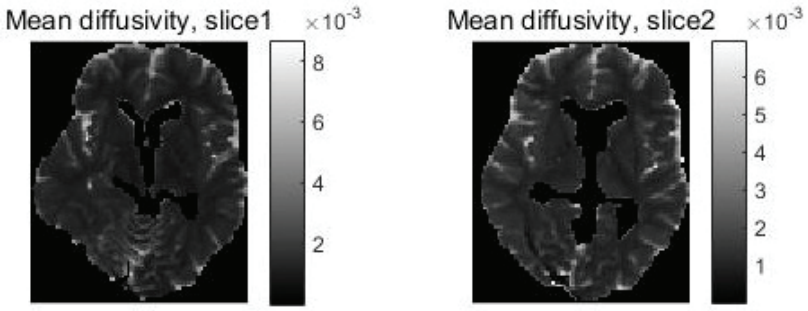
Control data. The data contain two subjects acquired respectively from a 46 year old and an 85 year old healthy Finnish male volunteers, where the second subject was from the same age group of the case data. In the acquisition protocol, we used all the combinations of 32 gradient directions with the b-values varying in the range 0, 62, 249, 560, 996, 1556, 2240 s/mm^2. After masking out the skull and the ventricles, the first dataset remains 18764 voxels and a total of 3 621 452 data points in the analysis.

The second subject has 21091 voxels with in total 4 070 563 data points to be analyzed. We calculated the data points by #voxel $\times$ (1 + # unique gradients $\times$ # unique b-values), where "1 " is from zero b-value and # refers to "the number of ". As an illustration, we plotted the parametric maps of estimated MD, MK and FA from the first two slices of both datasets shown in Figure 6 and 7. The number of voxels in slice 1 and 2 are 4537 and 4719, and are 4143 and 4573 from the first and the second subjects, respectively. About the first subject (the healthy volunteer with age 46), the estimates of MD are in [0, 8]$\times 10^{-3} mm^2/s$, MK estimates are in the range [0, 4] and the average estimated FA of all selected voxel is 0.2698. The second healthy volunteer has estimated values of MD are in [0, 6]$\times 10^{-3} mm^2/s$, of MK are in the range [0, 1.5]. The average of the estimated FA for all selected voxel is 0.2995. The color coded FA maps represent different orientations of the fibers in the brain. We use green, red and blue colors to depict three left-right, front-back and top-bottom orientations, respectively.

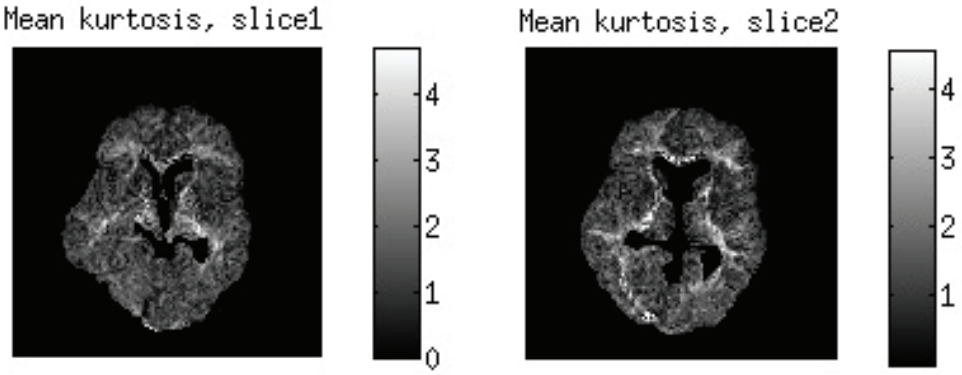
Case data. The dataset consists of diffusion-MR brain images of an 89 year old Finnish man diagnosed with LBD. After masking out the skull and the ventricles, we remain with a ROI containing 26104 voxels, with a total of 5 873 175 data points. For comparison, we also show the scalar-metric maps of the estimated MD, MK and FA from the first two slices, described in Fig. 8. The number of voxels in slice 1 and 2 are 4859 and 5212, respectively. The estimated values of MD are in [0, 4.7] $\times 10^{-3} mm^2/s$ and of MK is in the range $[0, 2]$. We corrected 34 voxels containing negative values of MD using the LS method and marked their spatial layout by the red stars shown in Fig. 8a. The average of estimated FA is 0.2883.

7. Conclusion and Discussion

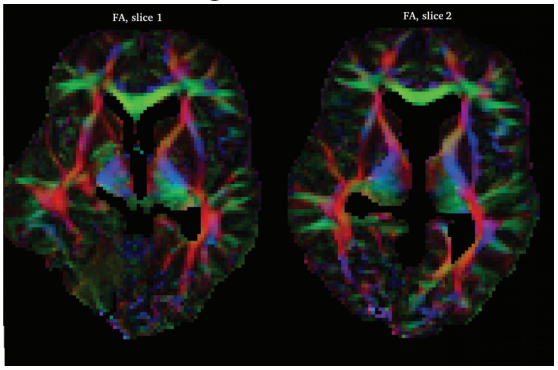
We are motivated by exploring possible ways to improve the diagnosis of one brain disorder, LBD, which so far is not clear in clinic. The proposed method in this paper provides a possible solution to estimate parameters with constrained DKI in diffusion MRI. The merits of the method is that can work for the data retrieved from low SNR. The method also considers the interactions between voxels, which hence can be used



(a) MD with regularization from DATA 1

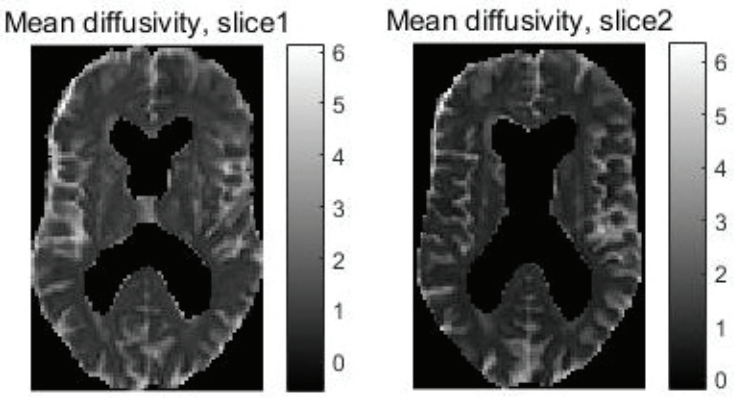


(b) MK with regularization from DATA 1

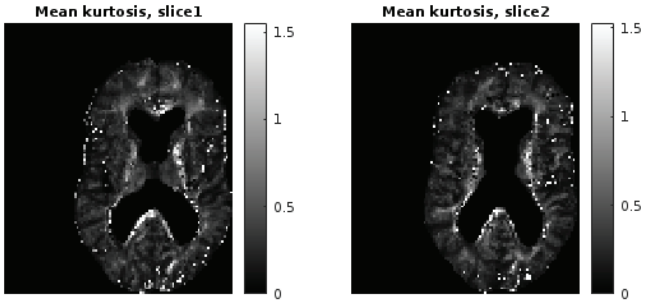


(c) FA with regularization from DATA 1

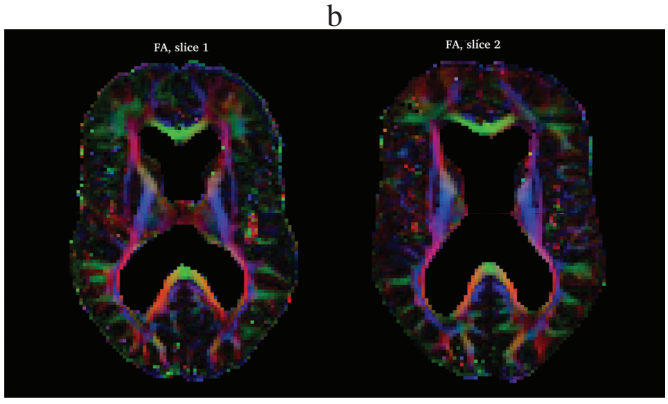
Figure 6: The MD, MK and FA maps from the first two slices of the healthy brain. The color coded FA maps also represent different fiber orientations in the brain. We use green color to indicate the left-right directions, the blue for the top-bottom and the red for the front-back. The color coded FA maps are obtained by using ExploreDTI Leemans et al. (2009) in the MATLAB environment.



a

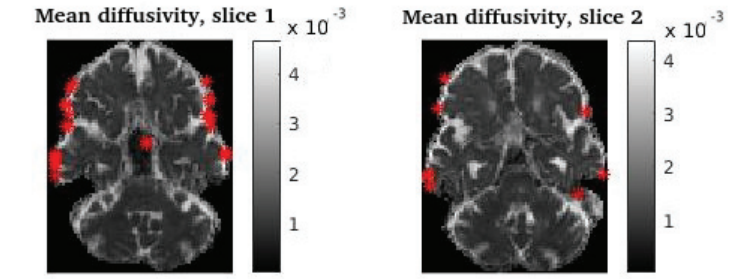


b

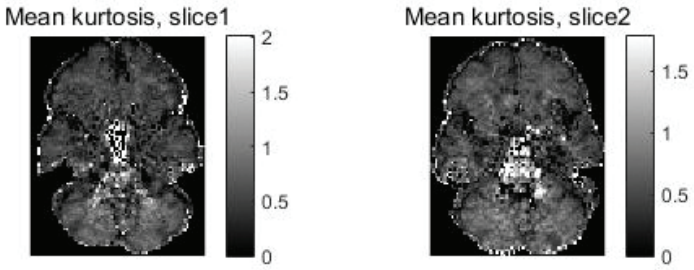


c

Figure 7: The MD, MK and FA maps from the first two slices of the healthy brain. The color code in the FA map describes the orientations of the fiber bundles: red, left-right; green, anterior-posterior; blue, superior-inferior. In Fig. 7a, the blue star describes the spatial distribution of the voxels having high values of MD which are far away from the visible region.

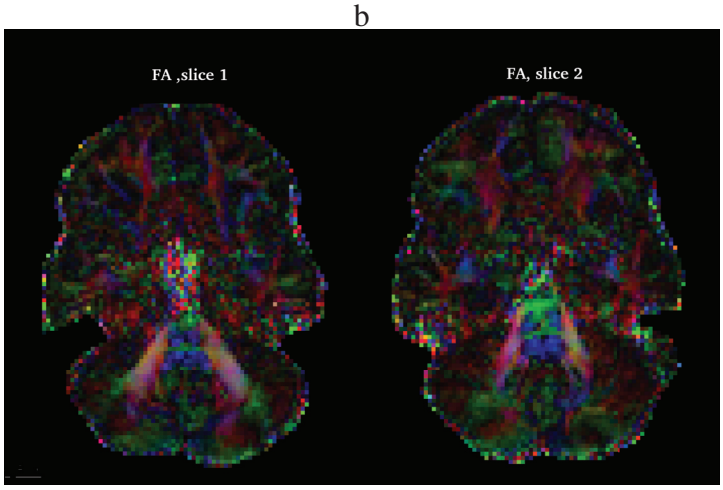


a



Regularized MK from case brain

b



c

Figure 8: MD, MK and FA maps of the first two slices from the case brain. The color code in the FA map represents the different orientations of the fibers in the brain. We use green color to indicate the left-right directions, the blue for the top-bottom and the red for the front-back. The color coded FA maps are obtained by using ExploreDTI Leemans et al. (2009) in the MATLAB environment. In Fig. 8a, the blue star describes the spatial distribution of the voxels having high MD values which are far away from the visible region.

in correction of noise level σ for image denoising and smoothing as illustrated in Fig. 2.

Studies on central nervous system and diagnosis of neurological diseases are important. DKI is an extension of traditional diffusion imaging techniques, characterizing both Gaussian and non-Gaussian probability distributions of water diffusion in vivo. Such a characteristic is capable of extracting independent and complementary information from complex structural tissues including cell barriers and relative compartments in different areas of the brain such as cortex and thalamus. DKI may be thought of a q-space formalism Ghosh et al. (2014), but it is naturally derived from the characteristic function of the signal decay under the Fourier transform by a truncated Taylor expansion. This well-defined quantity provides further insight on diffusional non-Gaussianity in terms of kurtosis. In this work we use Rician noise model to conduct the estimation with DKI in diffusion MRI. Using the correct noise model has the benefit of removing the bias, especially for the data retrieved in the low SNR regime compare with the common used solutions of weighted least squares. Using the state-of-the-art statistical methodology of data augmentation, we are able to work with generalized linear models when using the joint likelihood derived from the Rician density. The positivity constraints of kurtosis are imposed by the new parametrization in Section 2.3. Our second contribution is to describe a new computational method, the VB algorithm for the DKI estimation in the Bayesian framework and this method is more efficient than the Markov chain Monte Carlo method proposed in Gasbarra et al. (2014). It is now rather straightforward to extend the full scheme with other models of diffusion weighted signal decay.

The reason to use Bayesian modeling is that it provides many benefits in diffusion tensor estimation compared with the alternatives: we utilize the posterior distributions of all parameters of interest. This framework is capable of estimating not only the modes but also numerous statistics. In concreteness, from the tensor-variate posterior imposed by the prior, the proper specification of the tensor distribution, and conditioned on the given data, we are able to view the modes but also tensor moments and derived quantities, and to assess the uncertainty of all the model parameters. These statistics will help us to perform hypothesis testing for diffusion tensor-derived quantities in

the clinic practice and in material science. Furthermore, using Bayesian settings to interpret the kurtosis model, we are allowed to introduce and apply the variational approximation in the Bayesian analysis of DT-WRI data. The method will help us to reduce computation burden when more complicated signal models are encountered. In addition, the Bayesian framework allows us to model the interactions (or represent the dependence) between the tensors and the neighbouring simultaneously. The model should reduce the noise automatically by including information from the neighbours and provide averaging microstructural information on the different tissues of the brain, which may lead to significant applications in the clinic practice. In near term, we hope to provide such kind of detailed modeling and experimental results.

8. Acknowledgement

The authors would like to thank Emeritus Professor Antti Penttinen from the University of Jyväskylä who counseled them to use delta method and discussed them about stabilized Fisher scoring method, who reviewed the manuscript and made insightful comments, in which this work can be represented. Moreover, acknowledge Dr. Salme Kärkkäinen and Dr. Zitong Li for proofreading the manuscript. The authors are grateful to the CSC-IT Center for Science Ltd. for the use of their computer cluster. This work was funded by Doctoral Program in Computing and Mathematical Sciences (COMAS) and Department of Mathematics and Statistics, University of Jyväskylä.

Appendix

A. Gradient directions for real data

For each b-value, the MR-signal was measured in these 32 gradient directions.

u_x	u_y	u_z	u_x	u_y	u_z
-0.5000	-0.5000	-0.7071	0.7071	-0.4725	-0.5261
-0.5000	-0.5000	0.7071	-0.7071	-0.7071	-0.0002
0.7071	-0.7071	-0.0000	-0.7071	-0.4725	0.5261
-0.6533	-0.2706	-0.7071	0.7071	-0.4725	0.5261
-0.2087	-0.6756	-0.7071	0.4725	-0.7071	0.5261
0.0197	-0.7068	-0.7071	0.7071	-0.7071	0.0078
0.4212	-0.5679	-0.7071	-0.6364	-0.4252	0.6436
0.6899	-0.1549	-0.7071	-0.7060	-0.7060	0.0547
-0.6535	-0.2707	-0.7069	-0.2929	-0.7071	0.6436
-0.2929	-0.7071	-0.6436	0.2929	-0.7071	0.6436
0.2945	-0.7064	-0.6436	0.7071	-0.7071	0.0078
0.5150	-0.4861	-0.7061	0.7071	-0.2929	0.6436
0.7071	-0.2929	-0.6436	-0.7063	-0.7063	0.0489
-0.7071	-0.4725	-0.5261	0.0347	-0.7063	0.7071
-0.4725	-0.7071	-0.5261	0.7071	-0.7071	0.0115
0.5555	-0.6439	-0.5261	0.7071	0.0000	0.7071

B. Exponential moments of quadratic forms of Gaussian random variables

For $\psi \sim \mathcal{N}(\widehat{\psi}, \widehat{\Sigma})$, where $\widehat{\psi} \in \mathbb{R}^d$ and A is a $d \times d$ symmetric matrix,

$$
\begin{aligned}
&E\big(\exp\big(\psi^\top A\psi/2\big)\big)\\
&= (2\pi)^{-\frac{d}{2}} \det(\widehat{\Sigma})^{-\frac{1}{2}} \int_{\mathbb{R}^d} \exp\bigg(\frac{1}{2}\big\{(\psi^\top A\psi - (\psi-\widehat{\psi})^\top \widehat{\Sigma}^{-1}(\psi-\widehat{\psi})\big\}\bigg) d\psi\\
&= (2\pi)^{-\frac{d}{2}} \det(\widehat{\Sigma})^{-\frac{1}{2}} \exp\bigg(-\frac{\widehat{\psi}^\top \widehat{\Sigma}^{-1}\widehat{\psi}}{2}\bigg) \int_{\mathbb{R}^d} \exp\bigg(-\frac{\psi^\top(\widehat{\Sigma}^{-1}-A)\psi}{2}\bigg))\\
&\exp\big(\widehat{\psi}^\top \widehat{\Sigma}^{-1}\psi\big) d\psi\\
&= \det(\widehat{\Sigma})^{-\frac{1}{2}} \det(\widehat{\Sigma}^{-1}-A)^{-\frac{1}{2}} \exp\bigg(\frac{1}{2}\big\{\widehat{\psi}^\top \Sigma^{-1}(\widehat{\Sigma}^{-1}-A)^{-1}\widehat{\Sigma}^{-1}\widehat{\psi} - \widehat{\psi}^\top \widehat{\Sigma}^{-1}\widehat{\psi}\big\}\bigg)\\
&= \det(\mathrm{Id}-\widehat{\Sigma}A)^{-\frac{1}{2}} \exp\bigg(\frac{1}{2}\widehat{\psi}^\top \widehat{\Sigma}^{-1}\big\{(\mathrm{Id}-\widehat{\Sigma}A)^{-1}-\mathrm{Id}\big\}\widehat{\psi}\bigg),
\end{aligned}
$$

when $(\mathrm{Id}-\widehat{\Sigma}A)$ is positive definite, otherwise $E\big(\exp\big(\psi^\top A\psi/2\big)\big) = \infty$.

C. One-dimensional constrained Gaussian moments

Denote by $\Phi(t)$ and $\varphi(t)$ respectively the cumulative distribution function and the probability density of the standard Gaussian r.v. $G \sim \mathcal{N}(0,1)$. Consider the random variable $S := \mu + \eta G$ which is Gaussian $\mathcal{N}(\mu, \eta^2)$ distributed. Then

$$P(S>0) = P(\eta G + \mu > 0) = P(G > -\mu/\eta) = P(G < \mu/\eta) = \Phi(\mu/\eta)$$

due to the symmetry of the distribution of G.

For a test function $H(t)$ with weak derivative $h(t) = \frac{dH}{dt}(t)$ we have the Gaussian integration by parts formula

$$E\big(h(S)\big) = \eta^{-2} E\big(H(S)(S-\mu)\big). \tag{C.1}$$

Since $\phi'(x) = -x\phi(x)$ for the standard Gaussian density and $N(\mu, \eta^2)$ has density

$\eta^{-1}\phi\left(\frac{x-\mu}{\eta}\right)$, we have

$$E\big(h(S)\mathbf{1}(a<S<b)\big)=\eta^{-1}\int_a^b h(x)\phi\left(\frac{x-\mu}{\eta}\right)dx=$$
$$\eta^{-1}H(b)\phi\left(\frac{b-\mu}{\eta}\right)-\eta^{-1}H(a)\phi\left(\frac{a-\mu}{\eta}\right)-\eta^{-2}\int_a^b H(x)\frac{d}{dx}\left\{\phi\left(\frac{x-\mu}{\eta}\right)\right\}dx$$
$$=\eta^{-1}H(b)\phi\left(\frac{b-\mu}{\eta}\right)-\eta^{-1}H(a)\phi\left(\frac{a-\mu}{\eta}\right)-\eta^{-2}\int_a^b H(x)\phi'\left(\frac{x-\mu}{\eta}\right)\eta^{-1}dx$$
$$=\eta^{-1}H(b)\phi\left(\frac{b-\mu}{\eta}\right)-\eta^{-1}H(a)\phi\left(\frac{a-\mu}{\eta}\right)+\eta^{-1}\int_a^b H(x)\frac{x-\mu}{\eta^2}\phi\left(\frac{x-\mu}{\eta}\right)dx.$$

When $a\to-\infty$ and $b\to+\infty$ by taking limit $H(t)\phi\left(\frac{t-\mu}{\eta}\right)\to 0$ as $t\to\pm\infty$, we have

$$E\big(h(S)\big)=\eta^{-1}\int_a^b h(x)\phi\left(\frac{x-\mu}{\eta}\right)dx=$$
$$\eta^{-1}\int_{-\infty}^{\infty}H(x)\frac{x-\mu}{\eta^2}\phi\left(\frac{x-\mu}{\eta}\right)dx=\eta^{-2}E\big(H(S)(S-\mu)\big).$$

Consider now Eq. (C.1) with $H(s)=\mathbf{1}(s>0)$ and $h(s)=\delta_0(s)$, the Dirac δ-function with a point mass at 0,

$$\eta^2E\big(\delta_0(S)\big)=E\big(1(S>0)(S-\mu)\big)\Longleftrightarrow$$
$$\eta\phi(\mu/\eta)=E\big(1(S>0)S\big)-\mu\Phi(\mu/\eta).$$

Since

$$E\big(\delta_0(S)\big)=\frac{1}{\sqrt{2\pi\eta^2}}\int_{\mathbb{R}}\delta_0(x)\exp\left(-\frac{(x-\mu)^2}{2\eta^2}\right)dx=\eta^{-1}\phi(\mu/\eta),$$

we obtain

$$E\big(S|S>0\big)=\frac{E\big(\mathbf{1}(S>0)S\big)}{P(S>0)}=\mu+\eta\frac{\phi(\mu/\eta)}{\Phi(\mu/\eta)}.$$

Take now the test functions

$$H(s)=s^+=\max\{s,0\}=\mathbf{1}(s>0)s,\quad h(s)=\delta_0(s)s+\mathbf{1}(s>0).$$

It follows from Eq. (C.1) that

$$E\big(\delta_0(S)S\big)+P\big(S>0\big)=\eta^{-2}E\big(\mathbf{1}(S>0)S(S-\mu)\big)\Longleftrightarrow$$
$$0+\eta^2\Phi(\mu/\eta)+\mu E\big(\mathbf{1}(S>0)S\big)=E\big(\mathbf{1}(S>0)S^2\big)\Longleftrightarrow$$
$$E\big(S^2\big|S>0\big)=\frac{E\big(\mathbf{1}(S>0)S^2\big)}{P\big(S>0\big)}=\eta^2+\mu^2+\eta\mu\frac{\phi(\mu/\eta)}{\Phi(\mu/\eta)},$$

where

$$E\big(\delta_0(S)S\big) = \frac{1}{\sqrt{2\pi\eta^2}}\int_{\mathbb{R}} \delta_0(x)\exp\left(-\frac{(x-\mu)^2}{2\eta^2}\right)xdx = \eta^{-1}\phi(\mu/\eta)\times 0 = 0.$$

D. The KL divergence

The KL divergence between $q(\xi)$ and $p(\xi|y)$ is given by

$$\sum_{i=1}^{m}\int \log(\widehat{q}_i(\xi_i))\widehat{q}_i(\xi_i)d\xi_i - \int \log(p(\xi_1,\ldots,\xi_m|y))\prod_{i=1}^{m}\widehat{q}_i(\xi_i)d\xi_i,$$

which is non-increasing between consecutive VB steps.

In details,

$$\int \log(\widehat{q}(\theta))\widehat{q}(\theta)d\theta = -\frac{1}{2}\log|\Sigma_\theta| - 3\big(1+\log(2\pi)\big),$$
$$\int \log(\widehat{q}(\psi))\widehat{q}(\psi)d\psi = -\frac{1}{2}\log|\Sigma_\psi| - 9\big(1+\log(2\pi)\big),$$

$$\int \log(\widehat{q}(\sigma^2))\widehat{q}(\sigma^2)d\sigma^2 = \log(\hat{\nu}) + \log\Gamma(m) + m - (1+m)\frac{\Gamma'(m)}{\Gamma(m)},$$
$$\int \log(\widehat{q}(S_0))\widehat{q}(S_0)dS_0 = \frac{\widehat{\mu}}{2\widehat{\eta}}\frac{\phi\big(\widehat{\mu}/\widehat{\eta}\big)}{\Phi\big(\widehat{\mu}/\widehat{\eta}\big)} - \log(\widehat{\eta}) - \log\Phi\big(\widehat{\mu}/\widehat{\eta}\big) - \frac{1+\log(2\pi)}{2},$$
$$\int \log(\widehat{q}(\varphi_j))\widehat{q}(\varphi_j)d\varphi_j = \langle\cos\varphi_j\rangle\widehat{\kappa}_j = \frac{\widehat{\kappa}_j I_1(\widehat{\kappa}_j)}{I_0(\widehat{\kappa}_j)},$$

$$\int \log p(\theta,\psi,S_0,\sigma^2,\varphi|Y)\widehat{q}(\theta)\widehat{q}(\psi)\widehat{q}(\sigma^2)\widehat{q}(S_0)\prod_{j=1}^{m}\widehat{q}(\varphi_j)d\varphi_j dS_0 d\sigma^2 d\psi d\theta = \text{const}$$
$$+\langle\log\pi(S_0)\rangle - \frac{1}{2}\mu_\theta^\top\Omega\mu_\theta - \frac{1}{2}\text{Trace}(\Omega\Sigma_\theta) - \frac{1}{2}\text{Trace}\big(\mu_\psi^\top\Omega\mu_\psi\big) - \frac{1}{2}\text{Trace}(\Sigma_\psi:\Omega)$$

$$-(m+1)\langle\log(\sigma^2)\rangle - \frac{\langle S_0^2\rangle\langle\sigma^{-2}\rangle}{2}\sum_{j=1}^{m}\big\{Y_j^2 + \langle\exp\big(2Z_j\theta\big)\rangle\langle\exp\big(2\parallel Z_j\psi\parallel^2\big)\rangle\big\}$$
$$+\langle S_0\rangle\langle\sigma^{-2}\rangle\sum_{j=1}^{m}\langle\exp\big(Z_j\theta\big)\rangle\langle\exp\big(\parallel Z_j\psi\parallel^2\big)\rangle\langle\cos\varphi_j\rangle Y_j,$$

where

$$\langle\log(\sigma^2)\rangle = \log(\widehat{\nu}) - \frac{\Gamma'(m)}{\Gamma(m)}.$$

By putting the terms together, the KL divergence is expressed as

$$\text{const} + \frac{\widehat{\mu}}{2\widehat{\eta}} \frac{\phi\left(\widehat{\mu}/\widehat{\eta}\right)}{\Phi\left(\widehat{\mu}/\widehat{\eta}\right)} - \frac{\widehat{\mu}_0}{2\widehat{\eta}_0} \frac{\phi\left(\widehat{\mu}_0/\widehat{\eta}_0\right)}{\Phi\left(\widehat{\mu}_0/\widehat{\eta}_0\right)} + \log(\widehat{\eta}_0/\widehat{\eta}) + \log\Phi\left(\widehat{\mu}_0/\widehat{\eta}_0\right) - \log$$
$$\Phi\left(\widehat{\mu}/\widehat{\eta}\right) - \frac{1}{2}\log|\Sigma_\theta| - \frac{1}{2}\log|\Sigma_\psi| + \frac{1}{2}\mu_\theta^\top \Omega \mu_\theta + \frac{1}{2}\text{Trace}(\Omega\Sigma_\theta) + \frac{1}{2}\text{Trace}\left(\mu_\psi^\top \Omega \mu_\psi\right)$$
$$+ \frac{1}{2}\text{Trace}(\Sigma_\psi : \Omega) + \frac{\langle S_0^2\rangle\langle\sigma^{-2}\rangle}{2} \sum_{j=1}^{m} \left\{ Y_j^2 + \langle\exp\left(2Z_j\theta\right)\rangle\langle\exp\left(2 \parallel Z_j\psi \parallel^2\right)\rangle \right\}.$$

References

Andersen, A.H., 1996. On the Rician distribution of noisy MRI data. Magnetic resonance in medicine 36, 331–332.

Barmpoutis, A., Jian, B., Vemuri, B.C., Shepherd, T.M., 2007. Symmetric positive 4th order tensors & their estimation from diffusion weighted MRI. Information processing in medical imaging , 308–319.

Barmpoutis, A., Vemuri, B.C., 2010. A unified framework for estimating diffusion tensors of any order with symmetric positive-definite constraints, in: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, IEEE. pp. 1385–1388.

Barmpoutis, A., Zhuo, J., 2011. Diffusion Kurtosis Imaging: Robust estimation from DW-MRI using homogeneous polynomials, in: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, IEEE. pp. 262–265.

Basser, P.J., Mattiello, J., Le Bihan, D., 1994. Estimation of the effective self-diffusion tensor from the NMR spin echo. Journal of Magnetic Resonance, Series B 103, 247–254.

Basser, P.J., Mattiello, J., Turner, R., Le Bihan, D., 1993. Diffusion tensor echo-planar imaging of human brain, in: Proceedings of the SMRM.

Basser, P.J., Pajevic, S., 2003. A normal distribution for tensor-valued random variables: applications to diffusion tensor MRI. Medical Imaging, IEEE Transactions on 22, 785–794.

Berger, J.O., 2013. Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Casella, G., Berger, R.L., 2002. Statistical inference. volume 2. Duxbury Pacific Grove, CA.

Descoteaux, M., Deriche, R., Le Bihan, D., Mangin, J.F., Poupon, C., 2011. Multiple q-shell diffusion propagator imaging. Medical Image Analysis 15, 603–621.

Fisher, N.I., 1993. Statistical analysis of spherical data. Cambridge University Press.

Gasbarra, D., Liu, J., Railavo, J., 2014. Data augmentation in Rician noise model and Bayesian diffusion tensor imaging, in: arXiv preprint, arXiv 1403.5065.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. Bayesian data analysis. volume 2. Taylor & Francis.

Ghosh, A., Deriche, R., Moakher, M., 2009. Ternary quartic approach for positive 4th order diffusion tensors revisited, in: Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on, IEEE. pp. 618–621.

Ghosh, A., Milne, T., Deriche, R., 2014. Constrained diffusion kurtosis imaging using ternary quartics & MLE. Magnetic Resonance in Medicine 71, 1581–1591.

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

Helpern, J.A., Lo, C., Hu, C., Falangola, M., Rapalino, O., Jensen, J.H., 2009. Diffusional kurtosis imaging in acute human stroke, in: Proceedings 17th Scientific Meeting, International Society for Magnetic Resonance in Medicine, p. 3493.

Henkelman, R.M., 1985. Measurement of signal intensities in the presence of noise in MR images. Medical Physics 12, 232–233.

Hilbert, D., 1888. Über die Darstellung definiter Formen als Summe von Formenquadraten. Mathematische Annalen 32, 342–350.

Jaakkola, T.S., Jordan, M.I., 2000. Bayesian parameter estimation via variational methods. Statistics and Computing 10, 25–37.

Jensen, J.H., Helpern, J.A., 2010. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR in Biomedicine 23, 698–710.

Jensen, J.H., Helpern, J.A., Ramani, A., Lu, H., Kaczynski, K., 2005. Diffusional kurtosis imaging: The quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. Magnetic Resonance in Medicine 53, 1432–1440.

Kaipio, J., Somersalo, E., 2006. Statistical and computational inverse problems. volume 160. Springer Science & Business Media.

Koay, C.G., Basser, P.J., 2006. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. Journal of Magnetic Resonance 179, 317–322.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. The Annals of Mathematical Statistics , 79–86.

Leemans, A., Jeurissen, B., Sijbers, J., Jones, D., 2009. ExploreDTI: A graphical toolbox for processing, analyzing, and visualizing diffusion MR data, in: Proceedings of the International Society for Magnetic Resonance in Medicine, p. 3537.

Lindley, D.V., 1972. Bayesian Statistics: A review. SIAM.

Liu, J., 2015. An improved EM algorithm for solving MLE in constrained diffusion kurtosis imaging of human brain. arXiv preprint, arXiv 1507.06780 .

Mori, S., 2007. Introduction to diffusion tensor imaging. Elsevier.

Nadarajah, S., Kotz, S., 2008. Exact distribution of the max/min of two Gaussian random variables. Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 16, 210–212.

Ning, L., Laun, F., Gur, Y., DiBella, E.V., Deslauriers-Gauthier, S., Megherbi, T., Ghosh, A., Zucchelli, M., Menegaz, G., Fick, R., 2015. Sparse reconstruction chal-

lenge for diffusion MRI: Validation on a physical phantom to determine which acquisition scheme and analysis method to use? Medical Image Analysis 26, 316–331.

Ormerod, J.T., Wand, M., 2010. Explaining variational approximations. The American Statistician 64, 140–153.

Qi, L., Han, D., Wu, E.X., 2009. Principal invariants and inherent parameters of diffusion kurtosis tensors. Journal of Mathematical Analysis and Applications 349, 165–180.

Qi, L., Yu, G., Wu, E.X., 2010. Higher order positive semidefinite diffusion tensor imaging. SIAM Journal on Imaging Sciences 3, 416–433.

Šmídl, V., Quinn, A., 2006. The variational Bayes method in signal processing. Springer Science & Business Media.

Steven, A.J., Zhuo, J., Melhem, E.R., 2014. Diffusion Kurtosis Imaging: An emerging technique for evaluating the microstructural environment of the brain. American Journal of Roentgenology 202, W26–W33.

Tabesh, A., Jensen, J.H., Ardekani, B.A., Helpern, J.A., 2011. Estimation of tensors and tensor-derived measures in diffusional kurtosis imaging. Magnetic Resonance in Medicine 65, 823–836.

Tuch, D.S., 2002. Diffusion MRI of complex tissue structure. Ph.D. thesis. Citeseer.

Tuch, D.S., Weisskoff, R., Belliveau, J., Wedeen, V., 1999. High angular resolution diffusion imaging of the human brain, in: Proceedings of the 7th Annual Meeting of ISMRM, Philadelphia.

Veraart, J., Van Hecke, W., Sijbers, J., 2011. Constrained maximum likelihood estimation of the diffusion kurtosis tensor using a Rician noise model. Magnetic Resonance in Medicine 66, 678–686.

Zhu, H., Zhang, H., Ibrahim, J.G., Peterson, B.S., 2007. Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data. Journal of the American Statistical Association 102, 1085–1102.