

BAYESIAN ESTIMATION OF THE DIFFUSION TENSOR
FROM DIFFUSION WEIGHTED MRI DATA

ABSTRACT OF
A THESIS PRESENTED TO THE FACULTY
OF THE UNIVERSITY AT ALBANY, STATE UNIVERSITY OF NEW YORK
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
COLLEGE OF ARTS & SCIENCES
DEPARTMENT OF PHYSICS

TILMAN BIRNSTIEL

2007

BAYESIAN ESTIMATION OF THE DIFFUSION TENSOR FROM
DIFFUSION WEIGHTED MRI DATA

Tilman D Birnstiel

(Abstract)

Diffusion tensor imaging is a method of magnetic resonance imaging which allows observation of diffusion and thus makes it possible to characterize microscopic properties of biological tissue. First research on the scalar self-diffusivity D in isotropic media was done in 1965. Today it is well known that a scalar value is not sufficient for describing diffusion in non-isotropic media like biological tissue. A tensor formalism is needed to describe the directional dependency.

From the diffusion tensor, rotational invariants such as the trace or anisotropy indices are calculated to characterize abnormal diffusion in the brain due to diseases like stroke or schizophrenia. The diffusion tensor can also be used for investigation of fiber orientation in the brain (tractography).

Constrained and unconstrained linear and nonlinear least squares methods are commonly used to determine the diffusion tensor from diffusion weighted magnetic resonance images. This work reviews these methods and extends them by using Bayesian parameter estimation to incorporate the true noise characteristic of magnitude MR images which is given by the Rice distribution. The precision and accuracy of this new technique is compared to previous methods using simulated data sets.

The shape of the posterior is examined to determine which optimization algorithm is adequate for this problem. The Jacobian and the Hessian matrices of the posterior are derived. They can be used for optimization algorithms on the one hand and to derive error estimates for the best fit parameters on the other hand. The error estimates are tested on simulated data and shown to be valid.

BAYESIAN ESTIMATION OF THE DIFFUSION TENSOR
FROM DIFFUSION WEIGHTED MRI DATA

A THESIS PRESENTED TO THE FACULTY
OF THE UNIVERSITY AT ALBANY, STATE UNIVERSITY OF NEW YORK
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
COLLEGE OF ARTS & SCIENCES
DEPARTMENT OF PHYSICS

TILMAN BIRNSTIEL

2007

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

I would like to gratefully acknowledge the supervision of *Dr. Kevin H. Knuth*. Without his hints and ideas, this work had not been possible. I also would like to thank him for all the opportunities he offered me, such as a visit to an observatory, attending Maxent 2007 and many others which made my stay abroad even more interesting. I am very grateful to *Dr. Ariel Caticha*, *Dr. Akira Inomata* and again *Dr. Kevin Knuth* for their great lectures and the personal discussions from which I have benefited a lot. I would like to express my thanks to *Dr. Babak Ardekani* from the Nathan Kline Institute, who supported my work with ideas, papers and data. I also like to thank *Cheng Guan Koay* from the National Institute of Health for answering many questions about DTI and for sending me his code. I am thankful to all the people from the *International Student Services at Albany* and from the *Auslandsamt Würzburg* who made my stay abroad possible.

I also wish to thank *Anna Schüßler* for keeping me motivated during the last months.

I cannot end without thanking my parents, on whose constant encouragement and love I have relied throughout the last years. To them I dedicate this thesis.

Table of Contents

1	Introduction	1
2	Methods Of Magnetic Resonance Imaging	3
2.1	Measuring Spin Densities	3
2.1.1	Bloch Equation	3
2.1.2	Relaxation And Spin Echoes	4
2.1.3	Spatial localization	6
2.2	Diffusion Tensor Imaging	8
2.2.1	Fick's Law Of Diffusion	8
2.2.2	Diffusion Weighting	11
2.2.3	Anisotropic Diffusion	13
2.2.4	Applications Of DTI	16
2.3	Noise in MRI measurements	18
3	Parameter Estimation	24
3.1	Bayes' Theorem	24
3.2	Linearized Solutions	26
3.3	Cholesky Parametrization	29
3.4	Constrained Nonlinear Least Squares Method	31
3.5	Rician Likelihood Method	33
3.5.1	Rician Likelihood Function	33

3.5.2	Derivatives Of $\log P_R$	34
3.5.3	Approximations For $\ln(I_0(x))$	36
3.5.4	Error Estimates	39
4	Optimization Algorithms	43
4.1	The Nelder-Mead-Simplex Method	44
4.2	The Modified Full Newton Method	45
5	Results	49
5.1	Parameter Space	49
5.2	Diffusion Tensor Estimates	53
5.3	Error Estimates	56
5.4	Conclusions	57
	Bibliography	59
	A Marginalized Rice Distribution	64
	B Table of Acronyms	66

List of Figures

1	90°- and 180°-pulse	5
2	Echo train and decay times	7
3	Brownian Motion of a Particle	10
4	Stejskal-Tanner-Sequence	12
5	Diffusion-weighted images	15
6	Explanation of diffusion tensor anisotropy	16
7	Streamline plot following major eigenvectors	18
8	Integral transformation to polar coordinates	20
9	Rice distribution	21
10	Percent errors in approximations for $\ln(I_0(x))$	38
11	Percent errors in approximations of $I_1(x)/I_0(x)$ and $I_2(x)/I_0(x)$	39
12	Marginalized error bar	41
13	Simplex in two dimensional parameter space	45
14	Comparison of algorithms	48
15	Two dimensional slices through parameter space	51
16	Large scaled slice through parameter space	52
17	Estimates for the elements of the diffusion tensor.	54
18	Estimates for the traces of the diffusion tensor	55
19	Bias of the estimate versus SNR	55
20	Standard deviation of the estimates	55
21	Histogram of error estimates	57

Chapter 1

Introduction

Nuclear magnetic resonance (NMR) was first observed independently by Purcell and Bloch in 1946. Both were awarded the Nobel Prize in 1952 for their discoveries. It became an important method for the study of chemical compounds. It took until 1973 that NMR was used for imaging by Paul C. Lauterbur and Sir Peter Mansfield [1]. After their break-through work in this field, for which they received the Nobel Prize in 2003, and the introduction of fourier transform spectoscopy by Anderson and Ernst [2], this new imaging method developed quickly to become one of the most important methods in medical research and diagnosis. The experimental techniques have developed through the years – from a blurry picture of two tubes up to real-time images of the beating heart. The late 1990s saw the development of functional magnetic resonance imaging (fMRI) which allows imaging of brain activity.

NMR was also used for diffusion measurements since Erwin Hahn’s work on pulse techniques in 1950 [3]. Both methods, NMR diffusion measurements and MRI, were combined by Le Bihan et al. in 1985 and pushed forward by Basser et al. in the early 1990s [4]. Today, the combination of these methods is known as diffusion tensor imaging (DTI) and is used in many fields of medical and neurological research. It allows tracking fiber bundles in the brain (tractography), locating stroke or studying diseases such as multiple sclerosis or schizophrenia.

In the future these methods will be combined to allow further insights into brain connectivity and dynamic interactions.

The diffusion tensor mathematically describes the anisotropy of the diffusion. It can be used for further analysis like tractography. It is also needed to calculate anisotropy indices. The purpose of this work is to develop improved methods for diffusion tensor estimation, to compare them to other recent methods and to calculate confidence intervals of the estimated parameters.

Chapter 2

Methods Of Magnetic Resonance Imaging

2.1 Measuring Spin Densities

2.1.1 Bloch Equation

MRI uses the nuclear spin (commonly of Hydrogen protons) and its behavior under the application of external magnetic fields to obtain images of the biological tissue.

Without a magnetic field applied, the spins are randomly distributed resulting in a null net magnetization of the medium. Under the influence of an external magnetic field, nuclear spins precess around the field vector with an angular velocity of the Larmor frequency,

$$\omega_L = \gamma B. \tag{2.1}$$

The rotation axis of the proton spin is either aligned parallel or anti parallel to the field vector due to the quantization of the spin. Since the equilibrium distribution of the spin population is given by the Boltzmann distribution, and the parallel configuration is energetically favored, more spins are aligned parallel to the magnetic

field vector. For example, the ratio of parallel to antiparallel aligned hydrogen spins for a temperature of $T = 300$ K and a magnetic field of $B = 3$ T is

$$\frac{N_{\uparrow}}{N_{\downarrow}} = \text{Exp} \left[-\frac{\mu \mathbf{B}}{k_{\text{B}} T} \right] \approx 0.99998 . \quad (2.2)$$

The ratio of parallel and anti parallel spins is therefore still very close to unity, however there is a remaining net magnetization along the magnetic field.

The influence of an external field on the net magnetization of the spin ensemble can be derived from quantum mechanical expectation values for the spin components as shown in [5, 6]. Including phenomenological decay terms leads to the Bloch equation:

$$\frac{d}{dt} \langle \hat{\mathbf{s}} \rangle = -\frac{e}{m} \langle \hat{\mathbf{s}} \rangle \times \mathbf{B} + \begin{pmatrix} -\frac{1}{T_2^*} \langle \hat{s}_x \rangle \\ -\frac{1}{T_2^*} \langle \hat{s}_y \rangle \\ \frac{s_0 - \langle \hat{s}_z \rangle}{T_1} \end{pmatrix} \quad (2.3)$$

Application of a pulsed external field puts the spins into a nonequilibrium state. Both before and after the application of the external field, $\mathbf{B} = 0$ in the rotating frame. Equation (2.3) reduces to an exponential decay of the transversal magnetization (\perp z-axis) with a characteristic relaxation time T_2^* whereas the longitudinal magnetization (\parallel z-axis) follows a saturation curve up to s_0 with a characteristic time T_1 . These times depend on the environment of the spin, for example the cerebrospinal fluid has a different relaxation time than the gray matter in the brain.

The decay and saturation processes in Equation (2.3) were introduced phenomenologically, however the relaxation times can also be computed within the framework of a given formalism such as the Redfield theory (see [6, 7]).

2.1.2 Relaxation And Spin Echoes

By applying a transversal alternating magnetic field in resonance with ω_L , the net magnetization can be rotated away from parallel alignment with B_z . The angle of this

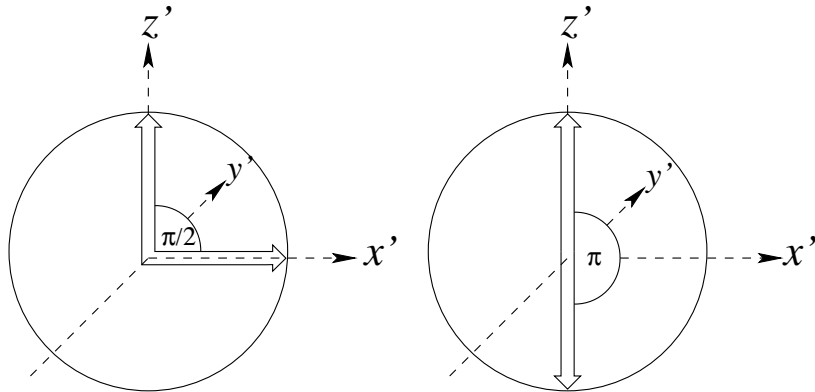


Figure 1: The net magnetization vector in a rotating reference frame. The $\pi/2$ -pulse flips the longitudinal magnetization into the $x' - y'$ plane, the π -pulse flips it into the $-z'$ -direction

rotation depends on the magnetic field and on its duration. In MRI this is achieved by a radio frequency (RF) pulse. The two most important pulses are the so called 90° - (or $\frac{\pi}{2}$ -) pulse and the 180° - (or π -) pulse. The names are derived from the angle by which the net magnetization is changed, as depicted in Fig. 1. The angle is determined by the duration and the strength of the pulse since the angular velocity is proportional to the B -Field.

If a 90° pulse is applied to a longitudinal magnetized ensemble, the net magnetization is flipped to the x - y -plane and rotates around the z -axis. The longitudinal magnetization has been turned into a transversal magnetization. The spins are rotating in phase with a angular velocity of $\omega_L = \gamma B_z$. However, due to microscopic inhomogeneities of the B -field and spin-spin interactions, the angular velocities of the spins are slightly different and the spins gradually dephase. This leads to a decay of the transversal magnetization of the sample. The signal decay due to the spin-spin interactions is tissue specific unlike the decay caused by field inhomogeneities. The influence of the latter can be minimized by a 180° pulse which is described below.

The longitudinal magnetization restores to its equilibrium state because of spin-lattice interactions. The energy absorbed from the 90° pulse is reemitted as RF radiation during this relaxation process. This signal can be recorded by induction

currents in surrounding coils.

If a 180° pulse is applied, Eq. (2.3) predicts that the spins are rotated by 180° around the axis of the applied pulse. Suppose the angle of the transversal magnetization in the x - y -plane is $\theta(t)$. After some time τ , one of the spins has dephased due to its higher angular velocity. The angle of the dephased spin is $\theta(t + \tau) + \delta(\tau)$.

Now if a 180° pulse is applied along the x -axis, the angle of the net magnetization is $-\theta(t)$ and the angle of the spin is $-\theta(t + \tau) - \delta(\tau)$. The advance of the spin toward the net magnetization has turned into an advance of the net magnetization. At the time 2τ the dephased spin will catch up with the net magnetization and the transverse magnetization is restored. This is called the spin echo.

However only dephasing processes due to field inhomogeneities disappear at 2τ , the so called spin-spin relaxation processes are not influenced. Thus, if this 180° pulse method is repeated, the height of each spin echo still decays, but the decay is only dependent on the tissue specific spin-spin interaction. The measured characteristic time is the pure T_2 time (the decay including the field inhomogeneity dephasing is called T_2^*). Both decay processes are depicted in Fig. 2.

2.1.3 Spatial localization

The previous methods allow one to measure spin intensities since the intensity of the collected spin echo is proportional to the number of RF emitting spins. However there is still no information about the spatial localization encoded in the signal. This can be achieved by applying different field gradients for *slice selection*, *phase encoding* and *frequency encoding*.

If the external magnetic field changes along the z -axis, then the Larmor frequency for a spin at z is given by $\omega_L = \gamma B(z)$. The position along the z -axis is now related to the frequency. A slice in the x - y plane with width Δz is described by a boxcar function in the space domain. The fourier transform of this results in a sinc function

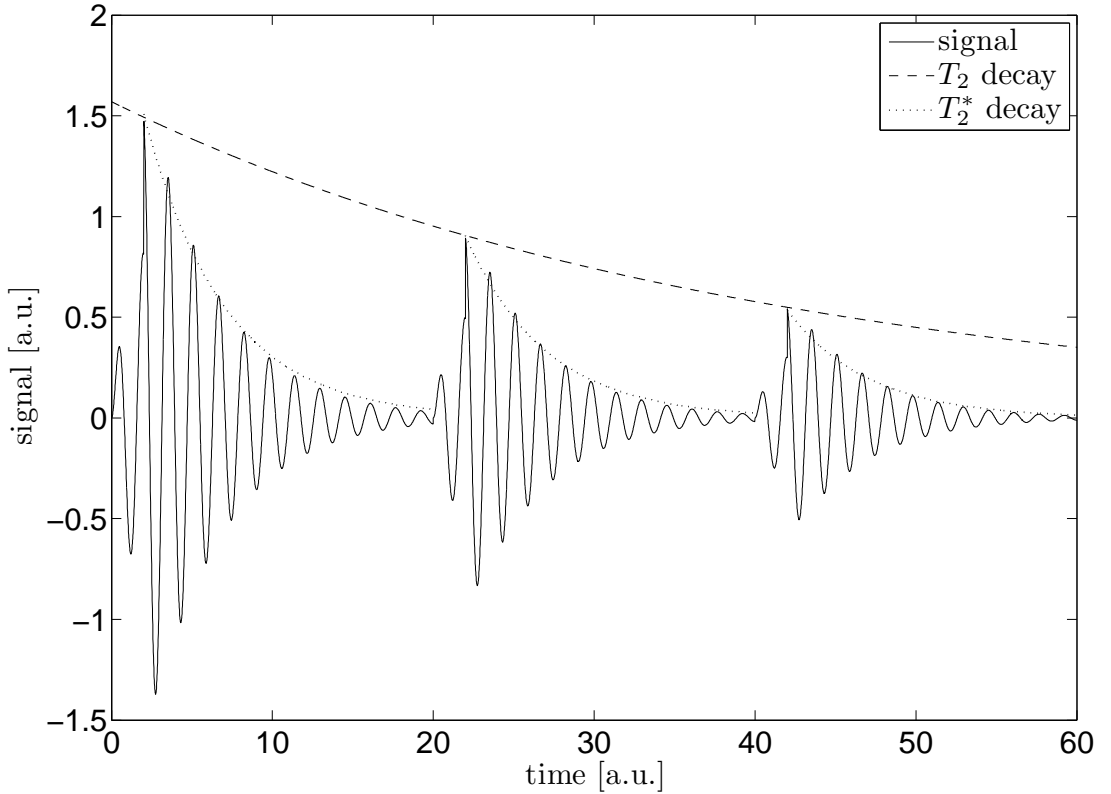


Figure 2: Echo trains are produced by repeated 180° pulses. The decay of each echo is due to inhomogeneities in the magnetic field and interactions. The decay from peak to peak is only caused by the interactions.

in the frequency domain. By applying an RF pulse with this sinc profile in frequency, one can excite spins in only this slice. This process is called *slice selection*. After the spins in one x - y plane have been excited the position of the spins within this plane must be encoded. This can be achieved by applying another field gradient in the y -direction. This gradient is only applied briefly resulting in a phase shift. For example, if the field strength at y_2 is higher than at y_1 then - according to Eq. (2.1) - the spins at y_2 precess with a higher Larmor frequency. Afterwards, all spins precess with the same frequency, but the phase shift of the spin within a row depends on the y -coordinate of the row. Similarly, the x -coordinate can be encoded with a gradient applied during the detection of the signal. This leads again to different precession frequencies along the direction of the gradient. The result is that one single spot in the x - y plane of the sample is being excited. The y -position of each spin is encoded

in the phase and the x -position is encoded in the frequency of the signal. The process is repeated for each slice of the sample and the spatial information is obtained by a two dimensional Fourier transformation.

2.2 Diffusion Tensor Imaging

Spin echoes have been used for diffusion measurements since the 1950s – long before magnetic resonance images were known. The most important work was done by Hahn [3], Torrey [8] and also Stejskal and Tanner [9]. In the 1980s, a method combining diffusion measurements and MRI was introduced, which gave rise to diffusion imaging [10]. In the beginning of this method, diffusion was only described by a scalar constant, but it became soon obvious that a tensor formalism is needed to describe the anisotropy of diffusion in biological tissue. The basic principles used in diffusion weighted imaging (DWI) and diffusion tensor imaging (DTI) will be discussed in the following section. A more comprehensive introduction to DTI can be found in [11, 12, 13] and [10].

2.2.1 Fick's Law Of Diffusion

The first mathematical description was given by *Adolf Fick* in 1855 [14] which is known as *Fick's Law of diffusion*. Fick's Law was only a phenomenological equation and as such is not based on any first principles. In 1860 *James Clerk Maxwell* used his kinetic theory of a mean free path to describe diffusion. However, Maxwell's derivation was not general since there exists no mean free path in liquids. In one of his famous 1905 papers [15], *Albert Einstein* introduced a general derivation, however this derivation relies on long and complicated arguments.

In 1989 Edwin Thompson Jaynes introduced an elegant derivation based on inferential principles [16]: The velocity of a gas particle is easily in the order of some hundred meter per second, but its magnitude and direction fluctuates wildly. The

diffusion process does not depend on this velocity, but rather on a mean velocity of the average drift movement which is given by

$$\bar{v} = \frac{x(t + \tau) - x(t - \tau)}{2 \tau} \quad (2.4)$$

where the time τ has to be much longer than the thermodynamical timescale.

The central limit theorem states that a sum of many i.i.d. variables with finite mean and standard deviation approaches the normal distribution. The Brownian motion of the particle is a result of a large number of small incremental movements due to collisions with the surrounding particles. Therefore, the probability that a particle moves from x to y is given by a Gaussian distribution:

$$P(y|x, I) = A \cdot \text{Exp} \left[-\frac{(y - x)^2}{2 \sigma^2(\tau)} \right] \quad (2.5)$$

The Gaussian distribution is symmetric. The best estimate for the next position of the particle is therefore the mean, $\bar{y} = x$, which yields a zero mean velocity. This misconception caused a lot of confusion. Diffusion is not a problem about the definite prediction derived from physical equations of motion but rather a problem of inference.

The additional information required to solve this problem is the current position of the particle, x . The position of the particle in the past and future will be denoted by z and y , respectively (see Fig. 3). The knowledge where the particle is now does not determine where it will be in the future (the probability distribution for this process is still the Gaussian centered at x) but it influences the inferences about the position in the past.

The probability distribution of the past position can be calculated with Bayes

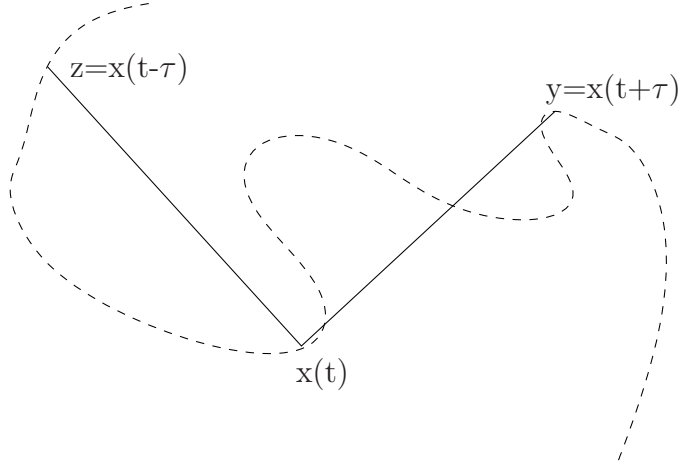


Figure 3: Random path of a particle and its positions in the past z , in the present x and in the future y .

theorem (which is described in detail in section 3.1) which is given by

$$P(z|x, I) = \frac{P(z|I) P(x|z, I)}{P(x|I)}. \quad (2.6)$$

The prior probability $P(z|I)$ is proportional to the particle density $n(z)$,

$$P(z|x, I) \propto n(z) \cdot \text{Exp} \left[-\frac{(z-x)^2}{2\sigma^2(\tau)} \right],$$

therefore

$$\log(P(z|x, I)) = \log(n(z)) - \frac{(z-x)^2}{2\sigma^2(\tau)}. \quad (2.7)$$

The most probable solution is found by maximizing Eq. (2.7):

$$\begin{aligned} \left. \frac{\partial \log P}{\partial z} \right|_{\hat{z}} &= 0 \\ \nabla_z \log(n(z)) &= \frac{\hat{z} - x}{\sigma^2} \\ \hat{z} &= x + \sigma^2 \nabla_z \log(n(z)) \end{aligned} \quad (2.8)$$

From Eq. (2.4) with \hat{z} and $\bar{y} = x$ we get the mean velocity,

$$\begin{aligned}\bar{v} &= \frac{x - (x + \sigma^2 \nabla_z \log(n(z)))}{2\tau} \\ \bar{v} &= -\frac{\sigma^2 \nabla(n)}{2\tau n(z)}.\end{aligned}\tag{2.9}$$

The average flux is therefore

$$\begin{aligned}J &= n \cdot \bar{v} \\ &= -\frac{\sigma^2}{2\tau} \cdot \nabla n.\end{aligned}\tag{2.10}$$

Letting $D = \frac{\sigma^2}{2\tau}$ yields Ficks law:

$$\mathbf{J} = -D \nabla n(x)\tag{2.11}$$

What Einstein accomplished in his eleven page paper, Jaynes accomplishes in a few lines.

2.2.2 Diffusion Weighting

The pulse sequence of an MRI experiment can be designed to be *diffusion weighted*. The most commonly used pulse sequence is called Stejskal-Tanner-Sequence [9] and is depicted in Fig. 4. If an additional gradient is applied along any direction, the precession frequency of the spins changes according to the change of field strength at their position. As the pulse is turned off, the spins precess with their former frequency but different phases.

If a spin is located at x_1 , and the magnetic pulse is applied in the x -direction for the duration δ , then the phase shift is

$$\phi_1 = \int_0^\delta \Delta\omega dt = \gamma G x_1 \delta,\tag{2.12}$$

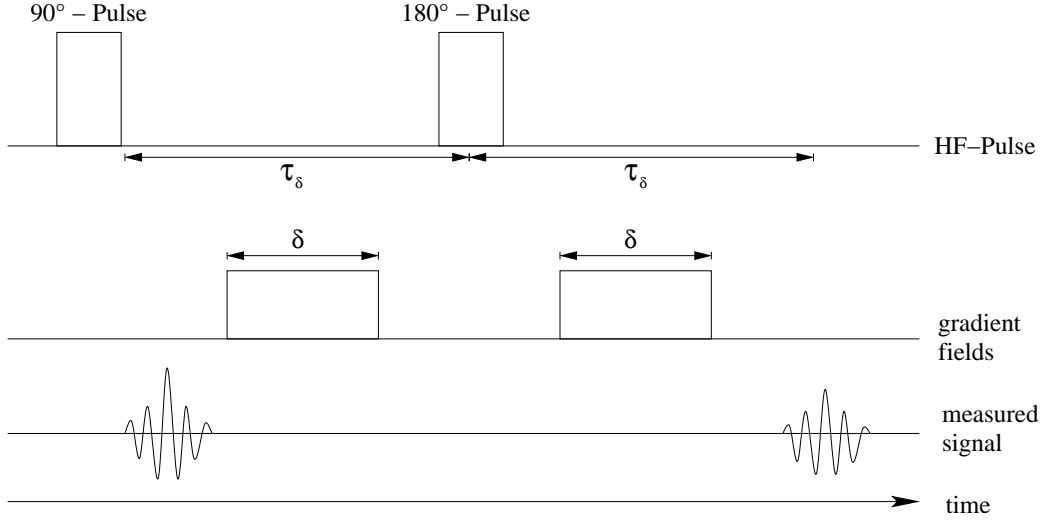


Figure 4: The Stejskal-Tanner-Sequence. The two gradient pulses before and after the 180° pulse cause the diffusion weighting. Spins diffusion along this gradient will have a different phase shift. This phase shift results in a decrease of net magnetization.

where $G = \Delta B / \Delta z$ is the gradient strength. After this dephasing gradient pulse, the 180° pulse flips the phase shift: $\phi_1 \rightarrow -\phi_1$. If a gradient pulse is applied the phase is shifted by

$$\phi_2 = \gamma G x_2 \delta. \quad (2.13)$$

This gradient pulse rephases the spins as long as they do not change their position ($x_1 = x_2$). If the position of the spin has changed due to diffusion then $\Delta\phi = \gamma G \delta (x_2 - x_1) \neq 0$ and the rephasing is not perfect. This results in a decrease of the net magnetization.

The net magnetization is a sum over all magnetic moments

$$M = M_0 \sum_{j=1}^N e^{i\phi_j} = M_0 \sum_{j=1}^N e^{i\gamma G \delta \Delta x_j}, \quad (2.14)$$

where M_0 is the transversal magnetization and $\Delta x = x_2 - x_1$ is the difference in the x -position. In the limit as $N \rightarrow \infty$, M can be computed as an expectation value. In DWI only the diffusion coefficient along one direction is measured, which reduces this to a one-dimensional problem. As mentioned in section 2.2.1, the probability distri-

bution for the one-dimensional Brownian motion of the spins is given by a Gaussian [17],

$$P(\Delta x|\tau) = \frac{1}{\sqrt{(4\pi\delta\tau)}} \text{Exp}\left[-\frac{\Delta x^2}{4D\tau}\right]. \quad (2.15)$$

The mean value of the magnetization is therefore

$$\begin{aligned} M &= M_0 \int_{-\infty}^{\infty} e^{i\gamma G\delta\Delta x} \frac{1}{\sqrt{(4\pi\delta\tau)}} e^{-\frac{\Delta x^2}{4D\tau}} dx \\ &= M_0 e^{-(\gamma G\delta)^2\tau D}. \end{aligned} \quad (2.16)$$

This is called the Stejskal-Tanner Equation and is commonly written as

$$M = M_0 e^{-bD} \quad (2.17)$$

where

$$b = (\gamma G\delta)^2\tau \quad (2.18)$$

is called the b -factor. This factor depends on the shape and duration of the pulse. Since a gradient pulse can not be perfectly rectangular shaped due to experimental constraints, a time integral of the gradient strength over the duration of the pulse has to be computed instead of Eq. (2.18). Further details can be found in [18] and [12]. Measurements with different b -factors along the same gradient direction are called *diffusion weighted images* (DWI).

2.2.3 Anisotropic Diffusion

In isotropic media, diffusion is described by Fick's Law, Eq. (2.11). If the medium is anisotropic, the diffusion coefficient depends on the direction. This directional

dependency is described by a tensor and Fick's law can be generalized to [4]:

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = - \begin{bmatrix} \mathcal{D}_{xx} & \mathcal{D}_{xy} & \mathcal{D}_{xz} \\ \mathcal{D}_{yx} & \mathcal{D}_{yy} & \mathcal{D}_{yz} \\ \mathcal{D}_{zx} & \mathcal{D}_{zy} & \mathcal{D}_{zz} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial C}{\partial x} \\ \frac{\partial C}{\partial y} \\ \frac{\partial C}{\partial z} \end{bmatrix} \quad (2.19)$$

Diffusion of uncharged particles is a symmetric process, hence the diffusion tensor is symmetric [19, 4]. However, the anisotropy of the medium leads to off-diagonal terms in the diffusion tensor \mathcal{D} . Isotropic diffusion can be described by a diagonal diffusion tensor with equal elements. The off-diagonal terms couple fluxes and concentration gradients in orthogonal directions. If the diffusion coefficient is replaced by a tensor, the b -factor in Eq. (2.17) is no longer only a scalar, it needs to contain coupling constants for each main axis direction of the tensor. In general, the tensor is not oriented along the reference frame of the MRI scanner, therefore off-diagonal coupling constants are needed as well. The b -factor becomes the \mathbf{b} -matrix and the Stejskal-Tanner equation is written as [20]

$$M = M_0 e^{-\sum_i \sum_j b_{ij} \mathcal{D}_{ij}} \quad (2.20)$$

The \mathbf{b} -matrix consists of products of the elements of the gradient directions

$$[\mathbf{b}]_{i,j} = b_i b_j \quad (2.21)$$

This enables one to write the Stejskal-Tanner Equation in matrix notation:

$$M = M_0 e^{-\mathbf{b} \cdot \mathbf{v}^T \cdot \mathcal{D} \cdot \mathbf{v}} \quad (2.22)$$

Here, b denotes the time integral over the gradient and determines the strength of the diffusion weighting, and \mathbf{v} is the unit vector in the direction of the diffusion weighting gradient pulse. Since the diffusion tensor is symmetric, it has six independent

parameters. Hence, at least six weighted measurements with non collinear directions are needed to determine the tensor completely.

Eq. (2.22) describes an exponential decay caused by diffusion. This means that the exponent has to be negative. Since the b -factor is always positive, the following condition must be true:

$$\mathbf{v}^T \cdot \mathcal{D} \cdot \mathbf{v} > 0 \quad (2.23)$$

According to [21] this is the definition of a positive definite matrix \mathcal{D} . This positive definite constraint on the diffusion tensor will be useful for the estimation of the tensor. For diffusion tensor imaging, diffusion weighted images are taken with different gradient directions and b -factors along with at least one unweighted image. A set of one unweighted and eight diffusion-weighted images¹ is shown in Fig. 5.

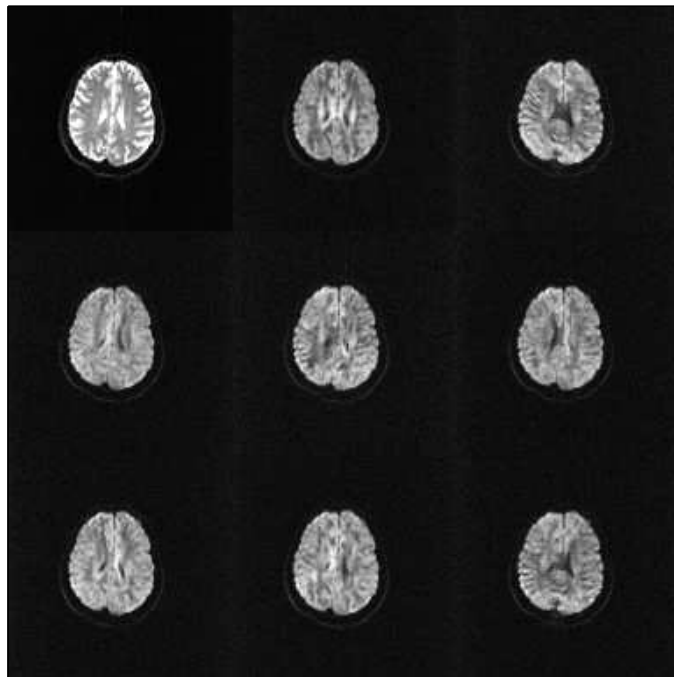


Figure 5: The top left image is an unweighted MRI image, the other eight images are diffusion-weighted along different gradient directions.

¹ Images are courtesy of Babak Ardekani, Center for Advanced Brain Imaging, Nathan Kline Institute, Orangeburg, New York

2.2.4 Applications Of DTI

Brain tissue is an highly anisotropic medium. The nerve cells are mainly positioned on the cortex of the brain (gray matter), whereas the axons (white matter), which act as the wiring, are mostly inside the brain. These connections are often bundled together into nerve cords. Diffusion inside and outside the axons is therefore constrained if its direction is perpendicular to the axis of the nerve cord. This is schematically shown in Fig. 6 This explanation is however not complete, the real cause of the diffusion is

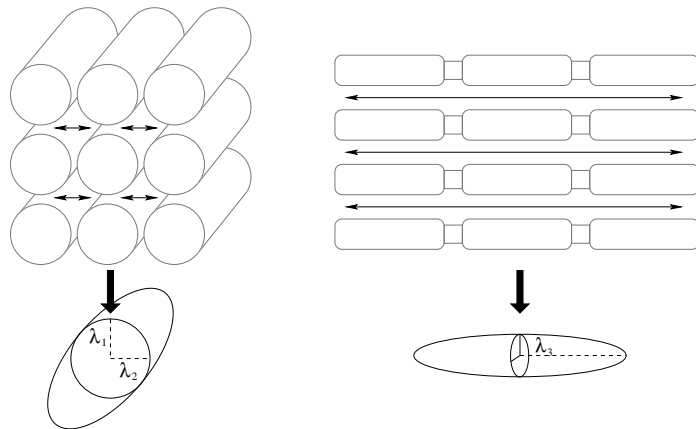


Figure 6: Schematic explanation of the anisotropy of diffusion inside the brain. The nerve cords suppress diffusion perpendicular to the axis of the bundle. The λ s denote the three eigenvalues of the diffusion tensor.

still subject of current research [22].

The fact that water molecules probe the microscopic structure of the tissue beyond the possible resolution of MRI and that diffusion is a new way to describe different properties of the tissue makes it such a powerful method. Some of the most important applications will briefly be mentioned in this section.

Stroke and traumatic injury

Usual MRI methods can detect stroke, but it usually takes about three hours until these methods become sensitive to stroke. DWI/DTI measurements show anomalous diffusion in the affected tissue already in the acute stage [23]. Reduced diffusion in

ischemic regions results in bright spots in diffusion weighted images. It is even possible to distinguish between early, subacute and late phase of stroke [10, 24]. Unlike DWI, DTI is also sensible to fiber-tract organizations such as axonal loss and incomplete remyelination following a stroke. The early detection of brain damage following traumatic injury can also be visualized with DWI/DTI. DTI allows investigation of the affected fiber tracts.

Surgical planning

Surgery on Brain tumors is very dangerous since nerv cords surrounding the tumor can easily be damaged. DTI is the only non-invasive method that can provide estimates of brain connectivity. It can differentiate between tumor tissue and surrounding fiber tracts. DTI can therefore be used for surgical planning in order to avoid damage to normal tissue surrounding a brain tumor [25].

Schizophrenia

The cause of schizophrenia is still unknown. DTI is used in this field of research since rotational invariants show anomalous behavior in some fiber tracts of diseased patients, especially in the corpus callosum which connects both hemispheres [26]. Other diseases that are believed or known to be related to brain white matter dysfunctions, such as multiple sclerosis or Alzheimer's disease are investigated this way as well [20].

Tractography

Following the main direction of the diffusion tensor (i.e. the eigenvector with the largest eigenvalue) makes is possible to infer the anatomical connections in the brain. Large fiber tracts can be detected this way, however crossing fibers are still a major problem. This method is a very recent topic in neurological research since it can be combined with functional MRI measurements to investigate brain connectivity [20].

A very simple streamline plot² produced from DTI data³ is shown in Fig. 7. It can be seen, that diffusion in the analyzed area is mainly directed along the sagittal axis.

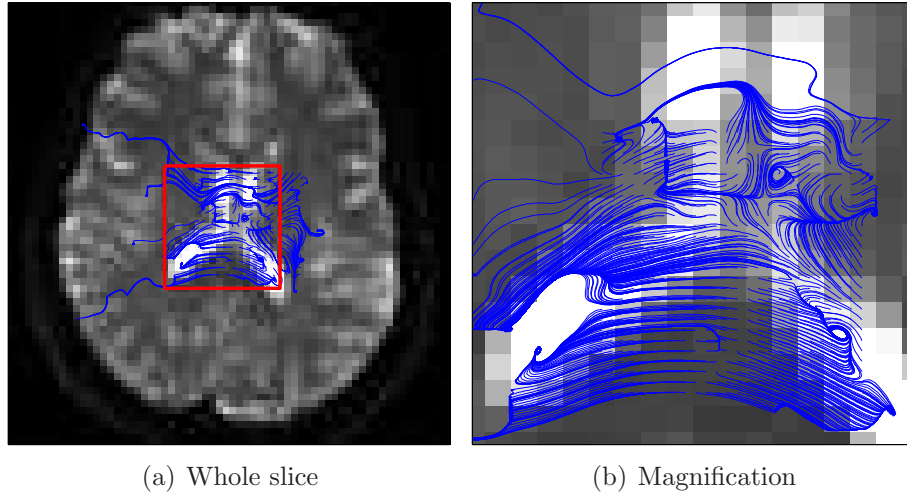


Figure 7: A simple 2D streamline plot with starting points in the region around the corpus callosum. The lines follow the eigenvectors with the largest eigenvalue.

2.3 Noise in MRI measurements

The noise in raw MRI data is assumed to be Gaussian with a standard deviation larger than the digital roundoff error which is therefore usually ignored [13]. The data is collected from the quadrature detectors and is processed by a Fourier transformation returning a complex and an imaginary part. The Gaussian characteristics of the signal will be preserved by this transformation since it is a linear and orthogonal transform [27].

The signal consists of a real and an imaginary part

$$A = A_R + i A_I \tag{2.24}$$

² The Streamlines are computed by the `streamline` function in Matlab[®], The Mathworks Inc., Natick, MA, USA

³ The DTI data for this plot is courtesy are courtesy of Babak Ardekani, Center for Advanced Brain Imaging, Nathan Kline Institute, Orangeburg, New York

Without loss of generality, the orientation of the coordinate system can be rotated so that the signal corresponds to the real axis [28]. Now, by adding the real (N_x) and imaginary (N_y) noise terms, Eq. (2.24) can be written as

$$A_N = A + N_x + i N_y = S \cdot e^{i \phi} \quad (2.25)$$

The image is usually derived by taking the magnitude of this complex signal. This is a nonlinear transform which does not preserve the Gaussian noise distribution:

$$S = \sqrt{(A + N_x)^2 + (N_y)^2} \quad (2.26)$$

The Gaussian distribution is given by

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.27)$$

The joint probability of the real and the imaginary noise is a simple product since both are independent

$$P(N_x, N_y|\sigma, A) = \frac{1}{2\pi\sigma^2} e^{-\frac{N_x^2 + N_y^2}{2\sigma^2}} \quad (2.28)$$

The distribution of the magnitude signal S can be derived by a change of coordinates from (N_x, N_y) to (S, ϕ) where:

$$\begin{aligned} x &= S \cos \phi - A \\ y &= S \sin \phi \end{aligned} \quad (2.29)$$

The Jacobian determinant of this coordinate transformation is given by:

$$\left| \frac{\partial(x, y)}{\partial(S, \phi)} \right| = \begin{vmatrix} \cos \phi & \sin \phi \\ -S \sin \phi & S \cos \phi \end{vmatrix} = S \quad (2.30)$$

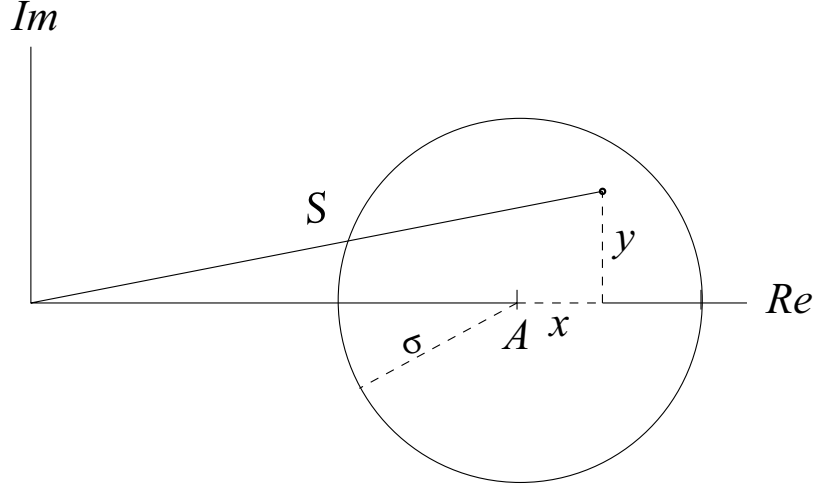


Figure 8: The integral is transformed to polar coordinates centered around $(A, 0)$

Using Eq. (2.30) and Eq. (2.29) in Eq. (2.28) yields

$$\begin{aligned}
 P(S, \phi | \sigma, A) &= \frac{S}{2 \pi \sigma^2} e^{-\frac{(S \cos \phi - A)^2 + (S \sin \phi)^2}{2 \sigma^2}} \\
 &= \frac{S}{2 \pi \sigma^2} e^{-\frac{S^2 + A^2}{2 \sigma^2}} e^{-\frac{S A \cos \phi}{\sigma^2}}
 \end{aligned} \tag{2.31}$$

The probability density function for the magnitude signal S is derived by marginalizing Eq. (2.31) over ϕ :

$$P(S | \sigma, A) = \frac{S}{2 \pi \sigma^2} e^{-\frac{S^2 + A^2}{2 \sigma^2}} \int_0^{2\pi} e^{-\frac{S A \cos \phi}{\sigma^2}} d\phi \tag{2.32}$$

The integral in Eq. (2.32) can be decomposed in two integrals over the intervals $[0, \pi]$ and $[\pi, 2\pi]$. Since the cosine is 2π -periodic, the second integration range can be shifted by -2π to $[-\pi, 0]$. Changing the direction of integration and substituting $\phi \rightarrow -\phi$, both yield a factor of -1 which cancel out. This way, the integral in Eq. (2.32) can be written as

$$\int_0^{2\pi} e^{-\frac{S A}{\sigma^2} \cos \phi} d\phi = 2\pi \int_0^\pi \frac{1}{\pi} e^{-\frac{S A}{\sigma^2} \cos \phi} d\phi \quad (2.33)$$

The remaining integral is an integral representation of the modified Bessel function of the first kind and order zero [29]. Therefore the probability density function becomes

$$P_{\text{Rice}}(S|\sigma, A) = \frac{S}{\sigma^2} \cdot e^{-\frac{S^2+A^2}{2\sigma^2}} \cdot I_0\left(\frac{S A}{\sigma^2}\right) \quad (2.34)$$

which is called the *Rice distribution* after Stephen O. Rice [30].

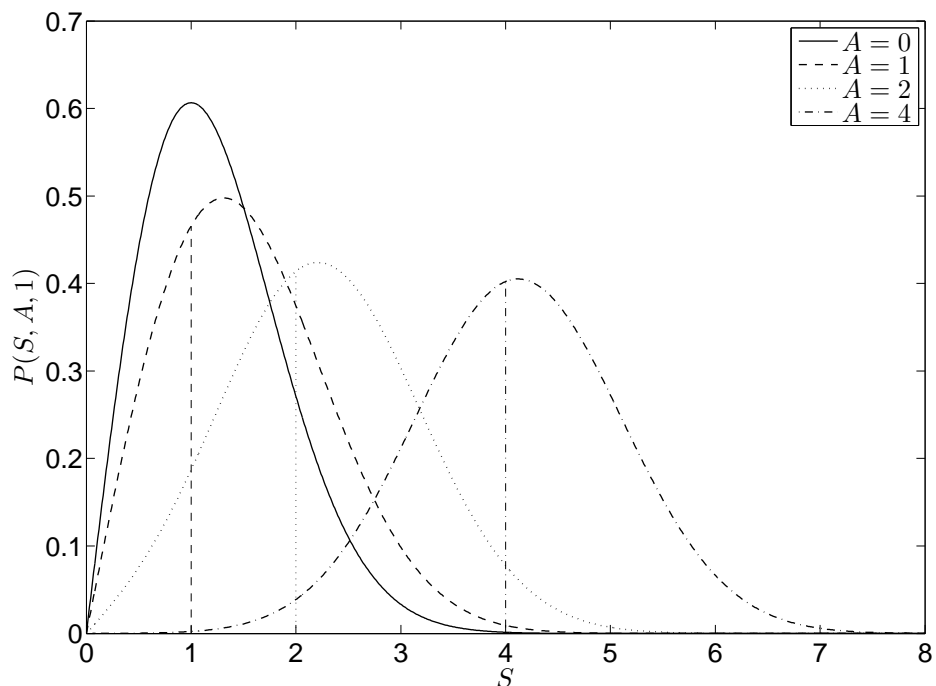


Figure 9: The Rice distribution for different signal to noise ratios (SNR). In these plots $\sigma = 1$ is fixed and the amplitude is varied. The vertical lines denote the position of A for the different curves. $A = 0$ coincides with the zero axis.

For signal to noise ratios greater than $A/\sigma \geq 3$, the Rice distribution starts to approximate the Gaussian distribution. The Rayleigh distribution is the limiting case for $A/\sigma = 0$ [27].

The σ in these measurements is unknown. One method to account for this is to marginalize the probability distribution over σ using a Jeffrey's non informative prior

[31],

$$P(\sigma|I) \propto \frac{1}{\sigma} \quad (2.35)$$

which is chosen because it is independent of the choice of scale.

In the case of a Gaussian distribution, this marginalization would give

$$\begin{aligned} P(\{x\}|\mu, I) &\propto \int_0^\infty \frac{1}{\sigma} \left(\frac{1}{\sqrt{2\pi} \sigma^2} \right)^N \text{Exp} \left[-\frac{\sum_{i=1}^N (x_i - \mu)^2}{2 \sigma^2} \right] d\sigma \\ &\propto (2\pi)^{-\frac{N}{2}} \int_0^\infty y^{N-1} \text{Exp} \left[-\frac{1}{2} \sum_{i=1}^N (x_i - \mu)^2 y^2 \right] dy \\ &\propto \frac{1}{2} \frac{1}{\sqrt{\pi^N}} \Gamma \left(\frac{N}{2} \right) \left[\sum_{i=1}^N (x_i - \mu)^2 \right]^{-\frac{N}{2}} \end{aligned} \quad (2.36)$$

which is called the *Student-t distribution* [32, 31].

The same approach was attempted with the Rice distribution. For the case of one measurement, the integral can be calculated, which is shown in appendix A. However, the resulting distribution is not normalizable. The marginalization over σ in the case of N measurements has not been completely solved yet.

Another approach is to estimate the σ from the data. According to [27] it can be estimated from regions where only noise is present by looking at the mean of the Rayleigh distribution since this is the limiting case for $\text{SNR} = 0$. The Rayleigh distribution is given by

$$P_R(S|\sigma) = \frac{S}{\sigma^2} e^{-\frac{S^2}{2\sigma^2}} \quad (2.37)$$

the mean and variance of which are

$$\begin{aligned} \langle S \rangle &= \sigma \sqrt{\frac{\pi}{2}} \\ \text{Var}(P_R) &= \left(2 - \frac{\pi}{2} \right) \sigma^2 \end{aligned} \quad (2.38)$$

Calculating the mean and variance of background data and comparing them to Eq.

(2.38) yields an estimate of the Gaussian sigma underlying the Rice distributed data.

Chapter 3

Parameter Estimation

3.1 Bayes' Theorem

Bayes' Theorem is named after, Reverend Thomas Bayes (1702–1761) who used it to compute inverse probabilities. This was found in his paper which was published by a friend after his death [33]. It is easily derived as a rewritten form of the product rule of probability which has been discovered before. Its generality was discovered by Laplace (1774) who was also the first to use it in a wide variety of problems of inference [34]. The following section shows how Bayes' Theorem can be used for parameter estimation of the diffusion tensor.

The product rule of probability is

$$P(M, D|I) = P(M|D, I) \cdot P(D|I) \tag{3.1}$$

where M and D are logical statements and I denotes all known prior information, the comma denotes *AND*. Since the *AND* is commutative, the product rule can be rewritten as

$$\begin{aligned} P(M, D|I) &= P(D, M|I) \\ P(M|D, I) \cdot P(D|I) &= P(D|M, I) \cdot P(M|I) \end{aligned} \tag{3.2}$$

Bayes theorem is derived by solving for one of the probabilities

$$P(M|D, I) = \frac{P(D|M, I) \cdot P(M|I)}{P(D|I)} \quad (3.3)$$

This equation becomes interesting for data analysis if D is associated with *data* and M with *model*. In this case, Eq. (3.3) gives the probability of the model (depending on some parameters) given the data. Maximizing this probability by adjusting the parameters of the model yields the set of parameters that has the highest probability given both the experimental data and one's prior info. To be able to do this, the different term on the right hand side of Eq. (3.3) need to be known.

$P(D|M, I)$ is called *likelihood*. It is the probability of measuring the data given that the hypothesized model is assumed to be true. It is often a description of the probability distribution of the noise. The likelihood can also represent uncertainty regarding the adequacy of the model. If it were a measurement without noise, this term would become a delta function (or the Kronecker delta for discrete values). In many cases this is a Gaussian distribution, in the case of magnitude MR data, the distribution is Rician.

$P(M|I)$ is called *prior probability*. It encodes the prior information about the parameters in a probability distribution. A maximum entropy distribution can be used to incorporate the effects of known constants (like moments of the distribution) and assume nothing beyond.

$P(D|I)$ is called *evidence*. It is a normalization factor since it does not depend on the model parameters. Since a constant does not change the position of a maximum, it can often be absorbed into a proportionality for parameter estimation. The evidence is useful for comparing the probabilities of the solutions of different models to select the most probable model.

As shown in [35], Bayes' rule is a special case of the method of Maximum relative Entropy (ME) for information in the form of data. Since the constraint of positive definiteness of the diffusion tensor can not be written in the form of moments, Bayes' theorem is used in this case. The constraint is implemented by hard coding it in the algorithm or by a special parametrization which yields only positive definite tensors.

3.2 Linearized Solutions

The Stejskal-Tanner equation is a transformable linear equation. This means that it can be transformed into a linear equation by taking the logarithm of both sides. However this does not conserve the noise distribution since it is not a linear transformation.

As already seen in Eq. (2.19), the diffusion tensor contains six independent parameters which will be denoted by

$$\mathcal{D} = \begin{bmatrix} D_1 & D_2 & D_3 \\ D_2 & D_4 & D_5 \\ D_3 & D_5 & D_6 \end{bmatrix} \quad (3.4)$$

To simplify the problem, the exponent in Eq. (2.22) is rewritten:

$$\begin{aligned}
-b \mathbf{v}^T \cdot \mathcal{D} \cdot \mathbf{v} &= -b (D_1 v_1^2 + 2 D_2 v_1 v_2 + D_4 v_2^2 + 2 D_3 v_1 v_3 + 2 D_5 v_2 v_3 + D_6 v_3^2) \\
&= \sum_{j=1}^6 W_j \beta_j
\end{aligned} \tag{3.5}$$

Where

$$W = -b (v_1^2, v_2^2, v_3^2, 2 v_1 v_2, 2 v_2 v_3, 2 v_1 v_3) \tag{3.6}$$

encodes the gradient vector design and

$$\beta = (D_1, D_4, D_6, D_2, D_5, D_3)^T \tag{3.7}$$

is the parameter vector.

Since more than one measurement is necessary to determine all six parameters, this notation can be extended. All gradient directions and the according weighting factors can be put into the W matrix which is therefore called the *gradient design matrix* (v_{ij} denotes the j th component of the i -th gradient vector if N different directions were used):

$$W = \begin{bmatrix} v_{11}^2 & v_{12}^2 & v_{13}^2 & 2 v_{11} v_{12} & 2 v_{12} v_{13} & 2 v_{11} v_{13} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{N1}^2 & v_{N2}^2 & v_{N3}^2 & 2 v_{N1} v_{N2} & 2 v_{N2} v_{N3} & 2 v_{N1} v_{N3} \end{bmatrix} \tag{3.8}$$

Using these notations, the logarithm of the Stejskal-Tanner equation, Eq. (2.22), can be rewritten and the signal measured along the i -th gradient direction is calculated as

$$\ln \left(\frac{M_i}{M_0} \right) = \sum_{j=1}^6 W_j \beta_j \tag{3.9}$$

This model can now be used for a linear least squares fit which minimizes the least

squares of the difference between the model and the data:

$$f_{\text{LLS}}(\boldsymbol{\beta}) = \sum_{i=1}^N \left(\ln \left(\frac{S_i}{S_0} \right) - \sum_{j=1}^6 W_{ij} \beta_j \right)^2 \quad (3.10)$$

Letting

$$\mathbf{y} = \left(\ln \left(\frac{S_1}{S_0} \right), \dots, \ln \left(\frac{S_N}{S_0} \right) \right)^T \quad (3.11)$$

simplifies the following calculation: The least squares estimator $\hat{\boldsymbol{\beta}}$ has to satisfy the following condition:

$$\nabla f_{\text{LLS}}(\boldsymbol{\beta}) \Big|_{\hat{\boldsymbol{\beta}}} \stackrel{!}{=} 0 \quad (3.12)$$

The derivative is carried out for a component:

$$\begin{aligned} \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_k} &= 2 \sum_{i=1}^N \left[\left(y_i - \sum_{j=1}^6 X_{ij} \hat{\beta}_j \right) \cdot (-1) \cdot \sum_{j=1}^6 X_{ij} \delta_{jk} \right] \\ &= -2 \sum_{i=1}^N \left[\left(y_i - \sum_{j=1}^6 X_{ij} \hat{\beta}_j \right) X_{i,k} \right] \stackrel{!}{=} 0 \end{aligned}$$

This equation is true for each component if

$$\mathbf{y} = \mathbf{W} \hat{\boldsymbol{\beta}} \quad (3.13)$$

Since \mathbf{W} is in general not a square matrix, Eq. (3.13) is solved for $\hat{\boldsymbol{\beta}}$ by the pseudoinverse \mathbf{W}^+

$$\hat{\boldsymbol{\beta}} = \mathbf{W}^+ \mathbf{y} \quad (3.14)$$

As in [36], this linear solution can be extended to a 7-parameter model in which the unweighted signal S_0 is unknown as well. In this case both the parameter vector and the gradient encoding matrix are extended:

$$\boldsymbol{\beta} = (\ln(S_0), D_1, D_4, D_6, D_2, D_5, D_3)^T \quad (3.15)$$

$$\mathbf{W} = \begin{bmatrix} 1 & -b_1 v_{11}^2 & -b_1 v_{12}^2 & -b_1 v_{13}^2 & -2b_1 v_{11} v_{12} & -2b_1 v_{12} v_{13} & -2b_1 v_{11} v_{13} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -b_N v_{N1}^2 & -b_N v_{N2}^2 & -b_N v_{N3}^2 & -2b_N v_{N1} v_{N2} & -2b_N v_{N2} v_{N3} & -2b_N v_{N1} v_{N3} \end{bmatrix} \quad (3.16)$$

The linearized solution gives a first estimate for the diffusion tensor, however it has to be noted that it has two major drawbacks:

- The noise is Rician distributed, so taking the logarithm of the data would skew the noise distribution. The least squares estimation is derived from a Gaussian likelihood which takes neither of these facts into account.
- The linearized solution does not take into account that the diffusion tensor is positive definite

However, the linearized solution is an analytic solution which can be solved very fast. It can be used as a starting point for more advanced estimation methods which are described later.

3.3 Cholesky Parametrization

As mentioned in Section 2.2.3, the diffusion tensor is known to be positive definite and symmetric. According to this, the trace of the tensor has to be greater than zero. Rotational invariants such as the trace or anisotropy indices are usually used in neuroscience to investigate abnormal diffusion in the brain [26].

The noise in DTI measurements can lead to results that are not in agreement with the positive definiteness constraint, so estimation methods are needed, which produce only positive definite answers.

The most common solution for this problem is the Cholesky decomposition [37, 38] which will be described in this section:

If a matrix \mathbf{R} is given by

$$\mathbf{R} = \begin{bmatrix} R_1 & R_2 & R_3 \\ 0 & R_4 & R_5 \\ 0 & 0 & R_6 \end{bmatrix} \quad (3.17)$$

where $R_1, R_4, R_6 > 0$ then the matrix product $\mathbf{R} \cdot \mathbf{R}^T$ is

$$\mathbf{R} \cdot \mathbf{R}^T = \begin{bmatrix} R_1^2 & R_1 R_2 & R_1 R_3 \\ R_1 R_2 & R_2^2 + R_4^2 & R_2 R_3 + R_4 R_5 \\ R_1 R_3 & R_2 R_3 + R_4 R_5 & R_3^2 + R_5^2 + R_6^2 \end{bmatrix} \quad (3.18)$$

According to [39], a symmetric matrix is positive definite if and only if all its principal minors are positive. The (i, j) -minor is the determinant of the matrix derived by removing the i -th row and the j -th column of a matrix. The principal minors are the (i, i) -minors. As can be easily seen in Eq. (3.18), this is always the case for the matrix $\mathbf{R}^T \cdot \mathbf{R}$. If an estimation method uses a diffusion tensor parameterized in the form $\mathcal{D} = \mathbf{R}^T \cdot \mathbf{R}$, then the solution will always be positive definite.

A decomposition from \mathcal{D} to \mathbf{R} will also be needed. The equation

$$\mathcal{D} = \mathbf{R}^T \cdot \mathbf{R} \quad (3.19)$$

can be solved for each component of \mathcal{D} which gives the following results:

$$\begin{aligned}
R_1 &= \pm\sqrt{D_1} \\
R_2 &= \frac{D_2}{R_1} \\
R_3 &= \frac{D_3}{R_1} \\
R_4 &= \pm\sqrt{D_4 - R_2^2} \\
R_5 &= \frac{D_5 - R_2R_3}{R_4} \\
R_6 &= \pm\sqrt{D_6 - R_3^2 - R_5^2}
\end{aligned} \tag{3.20}$$

This can be used to rewrite a positive definite, symmetric 3-by-3 matrix into the Cholesky parametrization. Since the linearized solution will be used as a starting point for advanced methods, it needs to be positive definite to be decomposed according to Eq. (3.20). Every real, symmetric matrix \mathbf{A} can be decomposed into

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{U}^T \tag{3.21}$$

where $\mathbf{U}^T \cdot \mathbf{U} = \mathbf{1}$ and \mathbf{D} is a diagonal matrix containing the eigenvalues of \mathbf{A} [38]. The eigenvalues can be changed to be positive (for example by taking the magnitude). The positive definite matrix \mathbf{A}_{PSD} is then obtained by

$$\mathbf{A}_{\text{PSD}} = \mathbf{U} \cdot \mathbf{D}' \cdot \mathbf{U}^T \tag{3.22}$$

3.4 Constrained Nonlinear Least Squares Method

This section describes the Constrained Nonlinear Least Squares (CNLS) Method after Koay et al. [36]. The basic idea is to perform a least squares fit, but unlike the linear solution, the unmodified data is used. The method of least squares is based on a Gaussian distribution of the noise. As described in section 3.1, the joint posterior

probability $P(M|D, I)$, which is given by

$$P(\{M_i\} | \{D_i\}, I) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(M_i - S_i)^2}{2\sigma^2}} \quad (3.23)$$

is to be maximized. Taking the logarithm of this probability distribution does not change the position of a maximum in the parameter space (only its height). Additionally, it simplifies the equations and has numerical advantages. Constant factors or summands can also be omitted. The logarithm of $P(\{M_i\} | \{D_i\}, I)$ shall be denoted by $\log P(\mathbf{R})$.

$$\log P(\mathbf{R}) = - \sum_{i=1}^N \left(S_i - \exp \left(\sum_{j=1}^7 W_{ij} \beta_j \right) \right)^2 \quad (3.24)$$

Here \mathbf{W} is the same gradient encoding matrix as in Eq. (3.16), the parameter vector is now rewritten in the Cholesky parametrization: the elements of the diffusion tensor \mathcal{D} in Eq. (3.15) are replaced by their Cholesky parameterized equivalents from Eq. (3.20). The standard deviation of the noise, σ is assumed to be equal for each measurement and can therefore also be omitted. Any unconstrained optimization algorithm can be used to maximize $\log P(\mathbf{R})$. The solution will always be positive definite since the constraint is implemented in the parametrization. Different optimization algorithms are discussed in section 4. The derivatives of $\log P(\mathbf{R})$, which can be used for optimization can be found in [36].

The results of the CNLS estimate have been shown to be biased to small trace values [36]. This is a result of the Gaussian likelihood function. The maximum of the Rice distribution is shifted to the right of the true amplitude. Using a Gaussian likelihood leads therefore to an overestimated signal which results in an underestimation of the diffusivity.

3.5 Rician Likelihood Method

This section discusses the Bayesian approach to constrained diffusion tensor estimation and represents the main innovation of this thesis. To avoid any bias as in the CNLS method, the true noise distribution – the Rice distribution – is used.

3.5.1 Rician Likelihood Function

The joint probability distribution of the Rice distributed data $\{S_i\}$, given the modeled signals $\{M_i\}$ is given by

$$P_{\text{Rice}}(\{S_i\}|\sigma, \{M_i\}) = \prod_{i=1}^N \frac{S_i}{\sigma^2} \cdot e^{-\frac{S_i^2 + M_i^2}{2\sigma^2}} \cdot I_0\left(\frac{S_i M_i}{\sigma^2}\right) \quad (3.25)$$

Again, the logarithm of this probability distribution is used, which shall be denoted by $\log P_{\text{R}}(M_i)$

$$\log P_{\text{R}}(M_i) = \sum_{i=1}^N \left\{ \ln\left(\frac{S_i}{\sigma^2}\right) - \frac{S_i^2 + M_i^2}{2\sigma^2} + \ln\left(I_0\left(\frac{S_i M_i}{\sigma^2}\right)\right) \right\}. \quad (3.26)$$

The constant summands can again be dropped, which gives the following log-likelihood

$$\log P_{\text{R}}(M_i) = \sum_{i=1}^N \left\{ -\frac{M_i^2}{2\sigma^2} + \ln\left(I_0\left(\frac{S_i M_i}{\sigma^2}\right)\right) \right\}. \quad (3.27)$$

The M_i are calculated by the Stejskal-Tanner equation, either in the normal tensor parametrization or in the Cholesky parametrization which leads to positive definite solutions. If positive definite solutions are to be achieved with the normal tensor parametrization, then the constraint has to be implemented by `if`-statements. All three principal minors of \mathcal{D} have to be tested to be positive definite. If any of these is non positive, then this set of parameters is to be rejected by returning $\log P(M_i) = -\infty$. This check slows down the algorithm and also leads to discontinuities in the $\log P$ -function.

3.5.2 Derivatives Of $\log P_R$

The derivatives of the $\log P_R$ -function are useful for gradient-ascent search algorithms as well as for error estimation. Since the derivatives of the function depend on the parametrization, first the derivatives of $\log P_R$ with respect of M_i are calculated. The derivatives of the Stejskal-Tanner equation with respect to the parameters are calculated afterwards for both parametrizations.

The first derivative of $\log P_R$ with respect to the model parameters is

$$\frac{\partial \log P_R}{\partial P_j} = \sum_{i=1}^N \left\{ -\frac{M_i}{\sigma^2} \partial_j M_i + \frac{1}{I_0 \left(\frac{M_i S_i}{\sigma^2} \right)} \cdot I_1 \left(\frac{M_i S_i}{\sigma^2} \right) \cdot \frac{S_i}{\sigma^2} \cdot \partial_j M_i \right\} \quad (3.28)$$

To simplify the notation, the derivative of $\log P_R$ with respect to the i -th model parameter is written as

$$\frac{\partial M}{\partial P_i} = \partial_i M. \quad (3.29)$$

The derivatives $\partial_j M_i$ denote the derivative of the i -th model signal with respect to the j th parameter (depending on the parametrization).

Differentiating a second time with respect to P_k yields:

$$\begin{aligned} \partial_j \partial_k \log P_R(M_i) = \sum_{i=1}^N \left\{ -\frac{\partial_k M_i \cdot \partial_j M_i}{\sigma^2} - \frac{M_i}{\sigma^2} \cdot \partial_{j,k} M_i \right. \\ \left. - \left(\frac{I_1 \left(\frac{M_i S_i}{\sigma^2} \right)}{I_0^2 \left(\frac{M_i S_i}{\sigma^2} \right)} \cdot \frac{S_i}{\sigma^2} \partial_k M_i \right) \cdot I_1 \left(\frac{M_i S_i}{\sigma^2} \right) \cdot \frac{S_i}{\sigma^2} \partial_j M_i \right. \\ \left. + \left[\frac{1}{2} \left(I_0 \left(\frac{M_i S_i}{\sigma^2} \right) + I_2 \left(\frac{M_i S_i}{\sigma^2} \right) \right) \cdot \frac{S_i}{\sigma^2} \cdot \partial_k M_i \right] \right. \\ \left. \cdot \frac{1}{I_0 \left(\frac{M_i S_i}{\sigma^2} \right)} \cdot \frac{S_i}{\sigma^2} \partial_j M_i + \frac{I_1 \left(\frac{M_i S_i}{\sigma^2} \right)}{I_0 \left(\frac{M_i S_i}{\sigma^2} \right)} \cdot \frac{S_i}{\sigma^2} \cdot \partial_{j,k} M_i \right\} \quad (3.30) \end{aligned}$$

Expanding and ordering by the derivatives of the model gives

$$\begin{aligned} \partial_j \partial_k \log P_{\text{R}}(M_i) = \sum_{i=1}^N \left\{ \partial_k M_i \cdot \partial_j M_i \left\{ -\frac{1}{\sigma^2} - \left(\frac{I_1 S_i}{I_0 \sigma^2} \right)^2 + \frac{1}{2} \left(\frac{S_i}{\sigma^2} \right)^2 + \frac{1}{2} \frac{I_2}{I_0} \left(\frac{S_i}{\sigma^2} \right)^2 \right\} + \right. \\ \left. \partial_{j,k} M_i \left\{ -\frac{M_i}{\sigma^2} + \frac{I_1 S_i}{I_0 \sigma^2} \right\} \right\} \end{aligned} \quad (3.31)$$

For simplicity, the arguments of the modified Bessel functions I_0, I_2 and I_0 have been omitted. In this equation, the derivatives of the model M_i with respect to the parameters are inserted.

Since the modeled signal M_i , calculated by the Stejskal-Tanner Equation, is an exponential decay, the derivatives are therefore especially easy to calculate. Since every derivative produces only an additional factor, it is also easy to implement these derivatives computationally.

The notation is chosen according to the one used in the algorithm: each parametrization has seven parameters, (P_1, \dots, P_7) . The first parameter is the unweighted signal M_0 , the other parameters are either the elements of the diffusion tensor ($P_2 = D_1, \dots$) in the non Cholesky parametrization, or the elements of the Cholesky decomposition of the diffusion tensor ($P_2 = R_1, \dots$). For the non Cholesky parametrization, the derivatives are:

$$\begin{aligned} \partial_1 M_i &= \frac{1}{M_0} \cdot M_i \\ \partial_2 M_i &= -b_i \cdot v_{i1}^2 \cdot M_i \\ \partial_3 M_i &= -b_i \cdot 2v_{i1}v_{i2} \cdot M_i \\ \partial_4 M_i &= -b_i \cdot 2v_{i1}v_{i3} \cdot M_i \\ \partial_5 M_i &= -b_i \cdot v_{i2}^2 \cdot M_i \\ \partial_6 M_i &= -b_i \cdot 2v_{i2}v_{i3} \cdot M_i \\ \partial_7 M_i &= -b_i \cdot v_{i3}^2 \cdot M_i \end{aligned} \quad (3.32)$$

The derivatives of the Cholesky parametrized model are:

$$\begin{aligned}
\partial_1 M_i &= \frac{1}{P_1} \cdot M_i \\
\partial_2 M_i &= -b_i \left(2P_2 v_{i1}^2 + 2P_3 v_{i1} v_{i2} + 2P_4 v_{i1} v_{i3} \right) \cdot M_i \\
\partial_3 M_i &= -b_i \left(2P_3 v_{i1}^2 + 2(P_2 + P_5) v_{i1} v_{i2} + 2P_3 v_{i2}^2 + 2(P_6 v_{i1} + P_4 v_{i2}) v_{i3} \right) \cdot M_i \\
\partial_4 M_i &= -b_i \left(2P_4 v_{i1}^2 + 2P_6 v_{i1} v_{i2} + 2(P_2 v_{i1} + P_7 v_{i1} + P_3 v_{i2}) v_{i3} + 2P_4 v_{i3}^2 \right) \cdot M_i \quad (3.33) \\
\partial_5 M_i &= -b_i \left(2P_3 v_{i1} v_{i2} + 2P_5 v_{i2}^2 + 2P_6 v_{i2} v_{i3} \right) \cdot M_i \\
\partial_6 M_i &= -b_i \left(2P_4 v_{i1} v_{i2} + 2P_6 v_{i2}^2 + 2(P_3 v_{i1} + (P_5 + P_7) v_{i2}) v_{i3} + 2P_6 v_{i3}^2 \right) \cdot M_i \\
\partial_7 M_i &= -b_i \left(2(P_4 v_{i1} + P_6 v_{i2}) v_{i3} - 2P_7 v_{i3}^2 \right) \cdot M_i
\end{aligned}$$

3.5.3 Approximations For $\ln(I_0(x))$

The arguments of the Bessel functions in Equations (3.27)-(3.31) are $M_i \cdot S_i / \sigma^2$. For high SNR, these functions grow exponentially. For a SNR of 15, I_0 is on the order of 10^{96} , for a SNR of 20 it explodes to 10^{172} . These values get scaled down since in all equations above, either the logarithm of the value is used or two Bessel functions are divided by each other. However this happens after the values of the Bessel functions are evaluated. Since these values easily become too big for normal machine precision, which in Matlab[®] is limited to $\approx 1.8 \times 10^{308}$, approximations for $\ln(I_0)$, I_1/I_0 and I_2/I_0 are needed.

A solution to this problem can be found in [40]. According to [29], the modified Bessel function of the first kind has an asymptotic behavior:

$$\begin{aligned}
I_\nu(x) \approx \frac{e^x}{\sqrt{2\pi x}} \left(1 - \frac{4\nu^2 + 1}{8x} + \frac{(4\nu^2 - 1)(4\nu^2 - 9)}{2! (8x)^2} \right. \\
\left. - \frac{(4\nu^2 - 1)(4\nu^2 - 9)(4\nu^2 - 25)}{3! (8x)^3} + \dots \right) \quad (3.34)
\end{aligned}$$

Therefore the logarithm of the modified Bessel function of the first kind and order

zero can be written as

$$\ln(I_0(x)) \approx x - \frac{1}{2} \ln(2\pi x) \cdot \ln\left(1 + \frac{1}{8x} + \frac{9}{2(8x)^2} + \frac{9 \cdot 25}{3!(8x)^3}\right) \quad (3.35)$$

To improve the speed of the algorithm, an approximation for small values can be used too. Using the series expansion from [29],

$$I_\nu(x) = \left(\frac{1}{2}x\right)^\nu \sum_{k=0}^{\infty} \frac{\left(\frac{1}{4}x^2\right)^k}{k!\Gamma(\nu + k + 1)} \quad (3.36)$$

gives for $\nu=0$

$$I_0(x) \approx 1 + \frac{x^2}{4} + \frac{x^4}{64} + \frac{x^6}{2304} + \frac{x^8}{147456} \quad (3.37)$$

The logarithm can be expanded to

$$\ln(x + 1) = \sum_{n=1}^{\infty} (-1)^n \frac{x^n}{n} \quad (3.38)$$

Using Eq. (3.37) in Eq. (3.38) gives

$$\ln(I_0(x)) \approx \frac{x^2}{4} - \frac{x^4}{64} + \frac{x^6}{576} - \frac{11x^8}{49152} + \mathcal{O}(x)^{10} \quad (3.39)$$

Similarly, approximations around $x = 1$ and $x = 2$ are derived. The intersection of the different approximations are computed numerically and used to define a piecewise-defined function approximating the $\ln(I_0)$ for the whole positive real axis. A plot of the different approximations is shown in Fig. 10

Further approximations need to be done for the fractions I_1/I_0 and I_2/I_0 . For large values, Eq. (3.34) can be used again. The exponential term e^x in Eq. (3.34) is independent of ν as is the root $\sqrt{2\pi x}$. These terms cancel out and the fractions are

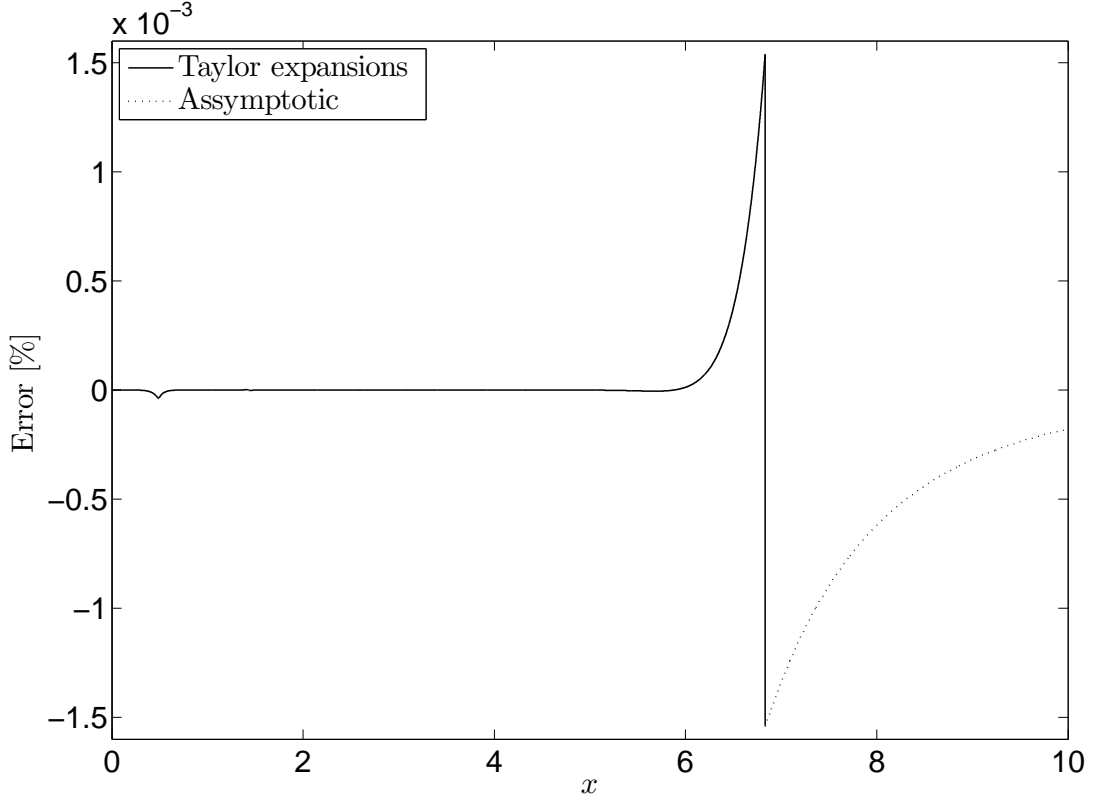


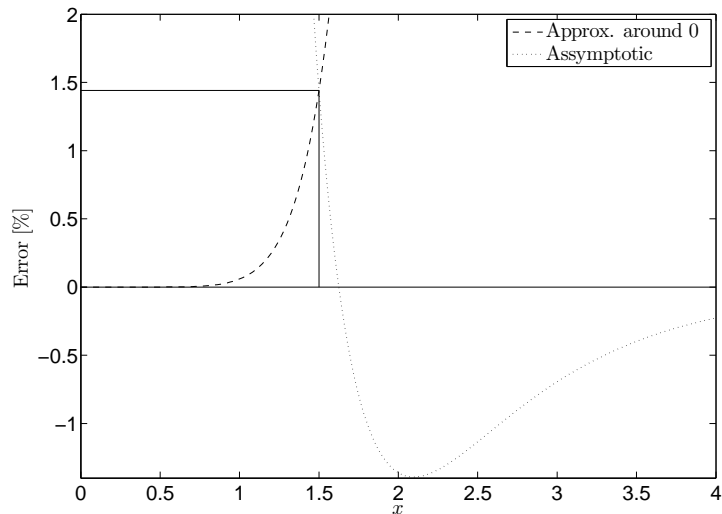
Figure 10: Percent errors in the approximations for the logarithm of the modified Bessel function of the first kind and order zero. Approximations around $x = 0, 1, 2$ are shown as well as the asymptotic behavior for large values. The maximum error of the piecewise defined function is 0.036 %.

therefore given by:

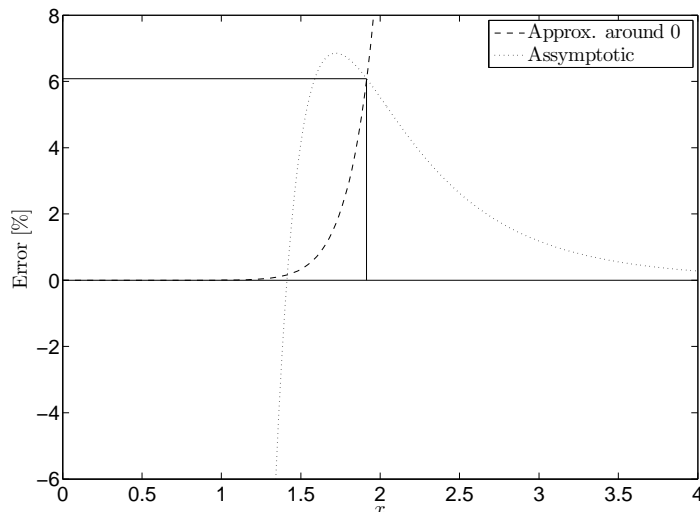
$$\frac{I_1(x)}{I_0(x)} \approx \frac{1 - \frac{3}{8x} + \frac{3 \cdot (-5)}{2!(8x)^2} - \frac{3 \cdot (-5) \cdot (-21)}{3!(8x)^3} + \dots}{1 + \frac{1}{8x} + \frac{9}{2!(8x)^2} + \frac{9 \cdot 25}{3!(8x)^3} + \dots} \quad (3.40)$$

$$\frac{I_2(x)}{I_0(x)} \approx \frac{1 - \frac{15}{8x} + \frac{15 \cdot 7}{2!(8x)^2} + \frac{15 \cdot 7 \cdot 9}{3!(8x)^3} + \dots}{1 + \frac{1}{8x} + \frac{9}{2!(8x)^2} + \frac{9 \cdot 25}{3!(8x)^3} + \dots} \quad (3.41)$$

For small values, a Taylor expansion around $x = 0$ is done. The approximation is again a piecewise defined function using the Taylor expansion up to the intersection and the asymptotic behavior after the intersection. It should be noted, that the asymptotic behavior is a very good approximation for large arguments, perceivable errors occur only for small arguments as shown in Figures 10 and 11.



(a) I_1/I_0



(b) I_2/I_0

Figure 11: Percent Errors in approximations for a) fraction of $I_1(x)/I_2(x)$ and b) fraction of $I_1(x)/I_2(x)$. The position and height of the maximum error is shown by the black line.

3.5.4 Error Estimates

To have a useful summary of the posterior probability, it is necessary to give not only the best estimate but also its reliability. The reliability depends on the peak surrounding the solution in the parameter space. A very narrow peak would result in small error bars which means that the estimate is very precise. A measure of reliability can be found using the second derivatives of the posterior probabilities [31]:

The best estimate is a maximum of the posterior obeying the following conditions:

$$\left. \frac{\partial \log P(\mathbf{x})}{\partial x_i} \right|_{\mathbf{x}_0} \stackrel{!}{=} 0 \quad (3.42)$$

where x_i are the parameters and \mathbf{x}_0 is the optimal solution.

Around this maximum, the posterior can be approximated by a multidimensional Taylor expansion:

$$L := \log P(\mathbf{x}_0) + \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \left. \frac{\partial^2 \log P}{\partial x_i \partial x_j} \right|_{\mathbf{x}_0} (x_i - x_{0i})(x_j - x_{0j}) \quad (3.43)$$

If the probability distribution is unimodal, it can now be written as

$$P(\mathbf{X}|\text{data}) \propto \exp\left(\frac{1}{2}(\mathbf{X} - \mathbf{X}_0)^T \cdot \nabla \nabla L(\mathbf{X}_0) \cdot (\mathbf{X} - \mathbf{X}_0)\right) \quad (3.44)$$

where $\mathbf{X} = [x_1, \dots, x_M]^T$ is the coordinate vector and $\mathbf{X}_0 = [x_{01}, \dots, x_{0M}]^T$ is the solution vector. This method is not applicable for multimodal probability distributions. The errors can now be estimated by calculating the standard deviation of this multivariate Gaussian. It is important to note that the standard deviation of the multivariate Gaussian along any coordinate axis need not be the right estimate since there can be correlations. This case is shown in Fig. 12 for the case of a two dimensional distribution. The correct way to infer the error bars for a multivariate Gaussian is described in [31]. The ij -th element of the covariance matrix is defined by

$$(\boldsymbol{\sigma})_{ij} = \langle (x_i - x_{0i})(x_j - x_{0j}) \rangle \quad (3.45)$$

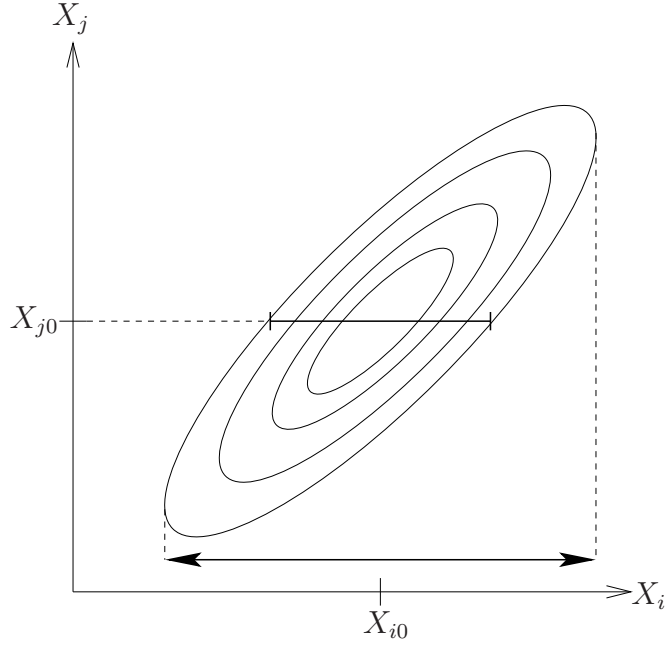


Figure 12: Marginalized error bar and best fit width along the according direction for the example of a two dimensional distribution.

Moving the coordinate center to \mathbf{X}_0 leads to the integral

$$\begin{aligned}
(\sigma)_{ij} &= \frac{1}{Z} \int_{-\infty}^{\infty} x_i x_j \exp\left(-\frac{1}{2} \mathbf{x}^T \cdot \nabla^2 L \cdot \mathbf{x}\right) d^N x \\
&= \frac{1}{Z} \int_{-\infty}^{\infty} x_i x_j \exp\left(-\frac{1}{2} \sum_{k=1}^N \sum_{l=1}^N x_k x_l (\nabla^2 L)_{kl}\right) d^N x \\
&= -2 \frac{\partial}{\partial (\nabla^2 L)_{ij}} (\ln(Z)) \\
&= \frac{\partial}{\partial (\nabla^2 L)_{ij}} \{\ln(\det(\nabla^2 L))\}
\end{aligned} \tag{3.46}$$

since

$$\begin{aligned}
Z &= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} \sum_{k=1}^N \sum_{l=1}^N x_k x_l (\nabla^2 L)_{ij}\right) d^N x \\
&= \frac{(2\pi)^{N/2}}{\sqrt{\det(\nabla^2 L)}}
\end{aligned} \tag{3.47}$$

is the normalization constant of a multivariate Gaussian distribution. The determi-

nant of a matrix \mathbf{A} can be calculated by

$$\det(\mathbf{A}) = \sum_j A_{ij} C_{ij} = \sum_j A_{ij} (-1)^{i+j} M_{ij} \quad (3.48)$$

where M_{ij} is the ij -th minor of A and C_{ij} is called cofactor. Differentiating a determinant with respect to A_{ij} is therefore in the case of $\mathbf{A} = \nabla^2 L$:

$$\frac{\partial}{\partial (\nabla^2 L)_{ij}} (\det(\nabla^2 L)) = C_{ij} \quad (3.49)$$

Using this in Eq. (3.46) gives the simple result

$$(\boldsymbol{\sigma}^2)_{ij} = \frac{C_{ij}}{\det(\nabla^2 L)_{ij}} \quad (3.50)$$

the right hand side is just the definition of the inverse of $(\nabla^2 L)_{ij}$ since it is a symmetric matrix. The covariance matrix therefore

$$\boldsymbol{\sigma}^2 = (\nabla^2 L)^{-1} \quad (3.51)$$

Hence, the errorbars for the i -th parameter can be calculated by taking the square root of the i -th diagonal element of the inverse of the second derivative matrix. The derivatives have been calculated in section 3.5.2.

Chapter 4

Optimization Algorithms

The previous chapter was concerned with finding and calculating cost functions which assign measures of preference to a given set of model parameters. The most probable set of model parameters is determined by the maximum of these functions. Unlike in the case of the linearized approach, the solution cannot be found analytically. Since the parameter space is seven dimensional, finding the true, global maximum can be very difficult. Sophisticated optimization algorithms are needed, which are able to find the global maximum of the cost functions.

As described earlier, the linear solution can be used as a first guess. As shown later, this starting point turns out to be close enough to the global maximum, so that more complex sampling methods are unnecessary. Sampling methods are very useful in finding the solution in large parameter spaces, but they are usually computationally intensive and slow. Since a single slice dataset of a DTI measurement contains usually 128×128 or 256×256 voxels, the speed of the algorithm is important.

The following sections introduces two different approaches of optimization algorithms, the *Nelder-Mead-Simplex method* using only function values and the *Modified Full Newton method* which also depends on derivatives of the log posterior probability.

4.1 The Nelder-Mead-Simplex Method

The Nelder-Mead-Simplex Method is a direct method, not depending on derivatives of the objective function, only on function values. It is a very reliable method for maximizing in bumpy parameter spaces, however it is not very efficient in the number of function evaluations [41].

Instead of one single point, which gets easily stuck in a local bump, a *surface* is used. A surface contains more information about the shape of the parameter space than a single point. This method uses a *simplex*, which is a geometrical figure consisting of $N+1$ points in N dimensions. In a two-dimensional parameter space, this would be a triangle (degeneracies such as three collinear points, are to be avoided). If this simplex is positioned near a peak, the simplex can climb up to the top by some simple rules:

Reflection: The worst point is selected and reflected through the centroid of the other N points. This step is shown in Figure 13. If the new found point is better than the old one, it is accepted. By reflection, the volume of the simplex is conserved and degeneracy is avoided.

Reflection with Expansion: If the new point is better than the old one (like above) and even better than the best point, then it seems valuable to explore further in this direction. Therefore the point is not only reflected but moved further in this direction.

Contraction: If the reflected point is worse than the original point, then it was moved too far in this direction. Hence the point between the centroid and the reflected point is evaluated. If this is not acceptable, the reflection of this point will be evaluated and used if it turns out to be better.

Shrinking: If all the above mentioned methods do not find an new acceptable point, then the top of the peak is probably within the simplex. To narrow the result down,

the simplex is contracted in each direction around the best point and the algorithm continues. The size of the simplex can be used for a termination rule. For example, if all sides of the simplex are smaller than a given tolerance value, the algorithm stops.

This methods is easy to implement, depends only on the function values, and is quite

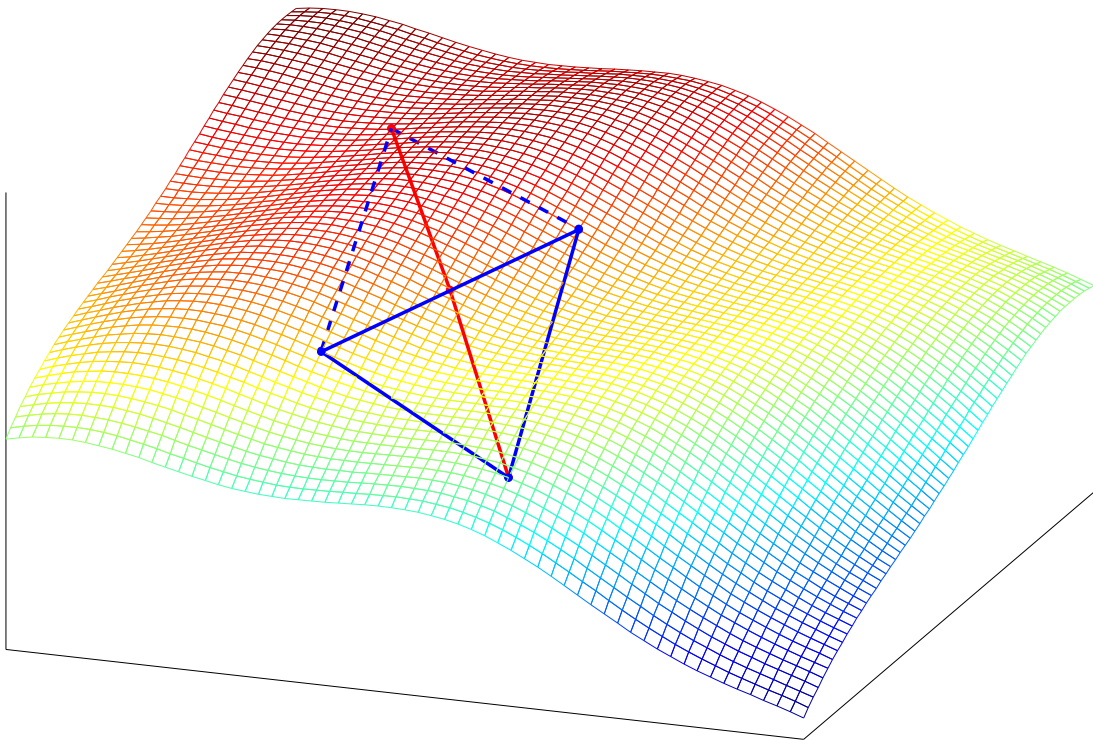


Figure 13: This shows one step of the simplex in a two dimensional parameter space. The worst point is reflected and accepted as a new point.

robust if the starting point is not too far from the maximum. It also can handle some types of discontinuities. The major drawback is that the large number of function evaluations makes it quite slow.

4.2 The Modified Full Newton Method

Some of the most commonly used optimization methods are the Newton, Gauss-Newton, Modified-Newton, and related methods. In this section, the main idea of

these methods will be discussed. We will then focus on the Modified Full Newton (MFN) Method as used in [36].

A one dimensional function can locally be approximated by a Taylor expansion

$$f(x + \delta) \approx f(x) + f'(x) \cdot \delta + \frac{1}{2}f''(x) \cdot \delta^2. \quad (4.1)$$

The best step size δ can be found by the zero of the derivative of Eq. (4.1) with respect to δ

$$\frac{d}{d\delta}f(x + \delta) \approx f'(x) + f''(x) \cdot \delta \stackrel{!}{=} 0, \quad (4.2)$$

the solution of which gives

$$\delta = -\frac{f'(x)}{f''(x)}. \quad (4.3)$$

Therefore the function value can be found by the iteration scheme [42]:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}. \quad (4.4)$$

The same approach can be used for optimization of multidimensional functions. According to [31], the Taylor expansion is given by

$$f(\mathbf{x} + \boldsymbol{\delta}) \approx f(\mathbf{x}) + \boldsymbol{\delta}^T \cdot \nabla f(\mathbf{x}) + \frac{1}{2} \boldsymbol{\delta}^T \cdot \nabla \nabla f(\mathbf{x}) \cdot \boldsymbol{\delta}. \quad (4.5)$$

The derivative with respect to $\boldsymbol{\delta}$ is

$$\nabla_{\boldsymbol{\delta}} f((x) + \boldsymbol{\delta}) = \nabla f(\mathbf{x}) + \nabla \nabla f(\mathbf{x}) \boldsymbol{\delta}, \quad (4.6)$$

which is set to zero and rewritten as

$$\mathcal{H}(x) \cdot \boldsymbol{\delta} = -\nabla f(\mathbf{x}). \quad (4.7)$$

Here $\mathcal{H}(x)$ denotes the Hessian matrix, which is composed of the second derivatives

of $f(\mathbf{x})$. An iteration method can again be used to obtain a solution:

$$x_{k+1} = x_k - \mathcal{H}^{-1}(x_k) \cdot \nabla f(\mathbf{x}_k) \quad (4.8)$$

This method is called the *Newton – Raphson* method. In practice, the normal Hessian matrix $\mathcal{H}(x)$ is not always used, but instead approximations are used in order to improve speed or robustness of the algorithm. One approach is called the *Gauss – Newton* method. In the case of *least squares* optimization, the Hessian consists of first and second derivatives of the residual function,

$$\mathcal{H}_{\text{LS}}(\mathbf{x}) = \mathbf{J}^T(\mathbf{x})\mathbf{J}(\mathbf{x}) + \sum_i r_i(\mathbf{x}) \nabla \nabla r_i(\mathbf{x}). \quad (4.9)$$

The second derivatives of the residual functions are dropped in the Gauss-Newton method, since they are small and computationally expensive. However, if these terms are not small, this method may fail to converge or converge slowly [42].

Another approach – called the *Levenberg-Marquardt* method – uses the Gauss-Newton Hessian and adds the identity matrix multiplied by a small positive constant to the Hessian. The constant is called the Levenberg-Marquardt parameter.

Adding constant terms to the diagonal of the Hessian does not change the directions of its eigenvectors, it only adds a constant to the eigenvalues:

$$(\mathcal{H}(\mathbf{x}) + c \mathbf{1}) \mathbf{e}_i = (\lambda_i + c) \mathbf{e}_i \quad (4.10)$$

The scale of the step size depends on the inverse of the Hessian as can be seen in Eq. (4.8). The inverse of a matrix is related to the reciprocal of its determinant [31]. Adding small diagonal elements to the Hessian prevents its determinant from getting too small – which would cause the step size to become too big.

Big steps are acceptable during the first iteration steps. As the algorithms approaches the peak the steps should become smaller. Hence the Levenberg-Marquardt

parameter should initially be set to zero and be adjusted during the iteration process. A recipe on how to do this can be found in [41].

The *Modified Full Newton* (MFN) method is identical to the Levenberg-Marquard method, except that it uses the full Hessian and not the simplified Gauss-Newton Hessian. For the estimation of the diffusion tensor with a Rician likelihood, the MFN and the Nelder-Mead method were compared. It was found that the MFN method was not significantly faster since the derivatives are computationally very intensive. The simplex method is therefore preferred due to its higher robustness.



Nelder Mead Algorithm	Modified Full Newton
<ul style="list-style-type: none"> • Many function evaluations but no second derivatives • More robust • Can handle some kinds of discontinuities 	<ul style="list-style-type: none"> • Very fast since derivatives for CNLS are simple • Linearized solution is close enough to the top of the peak • No discontinuities in posterior distribution
 Used for Rician method	 Used for CNLS method

Figure 14: Comparison of the Nelder-Mead simplex method and the MFN method

Chapter 5

Results

5.1 Parameter Space

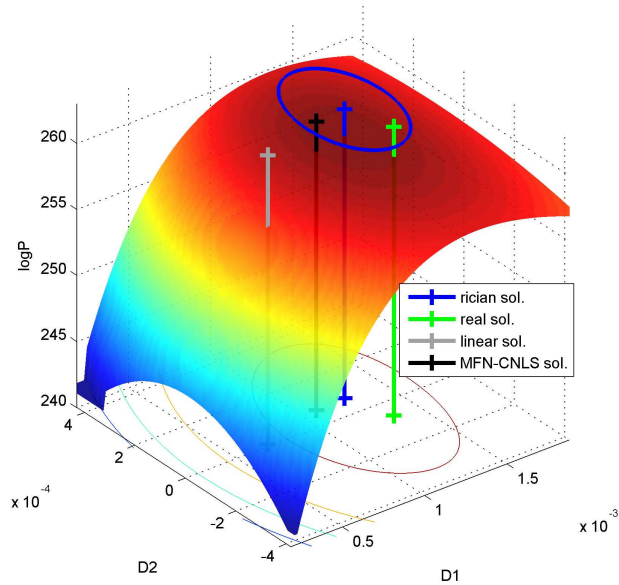
The choice of optimization algorithm depends on both the size and shape of the parameter space. In one or two dimensions it may still be possible to evaluate enough points in the parameter space to find the global solution. In high-dimensional parameter spaces the number of function evaluations would be too big to use this brute-force exhaustive-search approach. In these cases, sampling algorithms relying on Monte-Carlo methods are needed to explore the space and to find the global solution.

In the case of the estimation of the diffusion tensor, one can take advantage of the linearizable model. The linearized solution, as mentioned in section 3.2, has the big advantage of its speed. Since it can be calculated directly, its computation can be done within the order of 10^{-4} s on a 2 GHz Computer. This solution does not take the correct noise distribution into account but it is *close enough* to the real solution to be used as a starting point for further optimization methods. Close enough in this case means, that it is already on the correct peak in the seven-dimensional parameter space. This is shown in Fig. 15 a) and b), where both show a two-dimensional slice through the parameter space close to the true solution. It can be seen that the correct maximum is found by the algorithm. It is also obvious that both, the MFN-CNLS

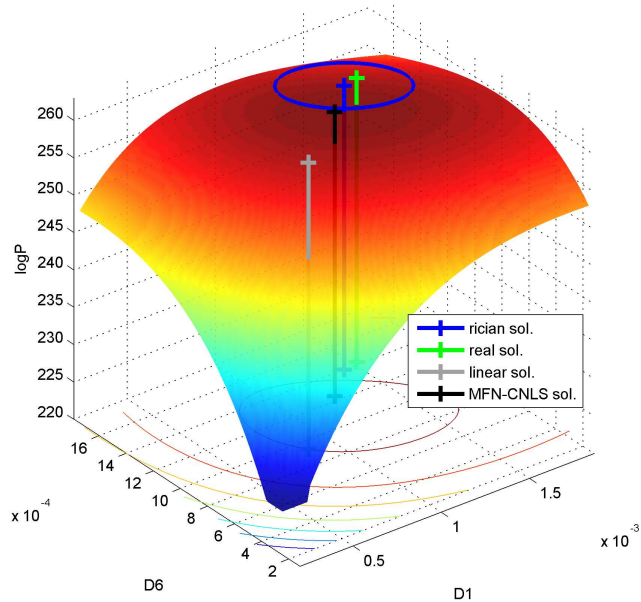
and the Rician method are closer to the true solution than the linear solution. The linear solution, however, needs not be positive definite. The linear solution shown in Fig. 15 is modified to be positive definite (as shown in section 3.3) and used as a starting point for the different non-linear methods. It should be noted that the surface shown in Fig. 15 is the space of the Rician posterior distribution, therefore the CNLS solution is not positioned on top of the peak.

Fig. 16 shows another two-dimensional slice through the parameter space. The uniformly dark blue colored area is forbidden since it would lead to a non positive-definite result. The peak is obviously smooth enough for Newton-type optimization methods. Sampling methods can therefore be avoided. However, Newton-type methods need first and possibly second derivatives of the log posterior probability which have been derived in section 3.5.2. These are computationally quite intensive which slows down the algorithm. Therefore, direct methods relying only on function values, which can be calculated faster, need not be slower.

However, the fact that the distribution is unimodal makes it possible to use the error estimation method discussed in section 3.5.4. The width of the peak, which is a measure of the uncertainty of the solution, is estimated from the second derivatives of the log posterior probability at the maximum.



(a) Slice along D1-D2 plane



(b) Slice along D1-D6 plane

Figure 15: Example of two slices through the Rician parameter space for a dataset with $\text{SNR} = 5$. The positions of the different solutions are marked with lines. The blue ellipse shows the error estimates for the Rician estimate.

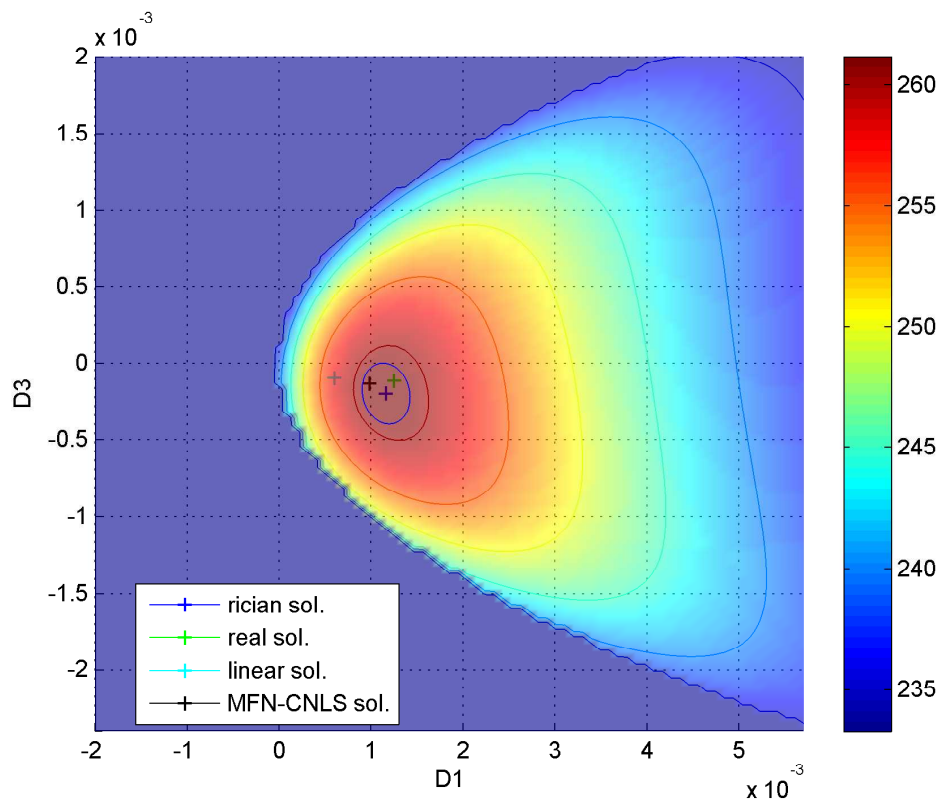


Figure 16: Example of a slice through the Rician parameter space with a wide field of view. The dark blue area is forbidden by the positive definiteness constraint.

5.2 Diffusion Tensor Estimates

Several typical tensors were tested with different SNRs. The linearized, the MFN-CNLS, and the introduced Rician probability method were compared. To obtain results that can be compared to [36] the same 23 gradient vectors (after [43]) and b -factors were used. For each gradient direction the b -factors 50, 500 and 1000 were used resulting in 69 data points for each data set. The unweighted signal is assumed to be 1000 arbitrary units, and the true signal is computed with these vectors and b -factors using Eq. (2.22). The noise can then be simulated by the following formula

$$S_i = \sqrt{(A_i + N(0, \sigma))^2 + N(0, \sigma)^2} \quad (5.1)$$

resulting in Rician distributed data. Here A_i denotes the true signal amplitude from Eq. (2.22). For each tensor, 10 000 measurements were simulated and used to test both algorithms. Typical results for the entries of the diffusion tensor are shown in Fig. 17. It can be seen, that the accuracy of the MFN-CNLS and the Rician method is similar for the off-diagonal terms (D_2, D_3, D_5), however the width of the MFN-CNLS estimates is slightly smaller. The true solution is marked with a vertical line from which can be seen, that the MFN-CNLS method is inaccurate for the diagonal terms (D_1, D_4, D_6). This results in a bias of the trace estimate which is shown for SNR = 5 and SNR = 15 in Fig. 18. The bias and width of the estimates has been measured for both methods and different signal to noise ratios as shown in Fig. 19. The Rician estimate shows a small positive bias here. This bias seems to depend on the chosen tensor. It was tested for several tensors and unlike the MFN-CNLS bias its does not generally over- or underestimate the trace.

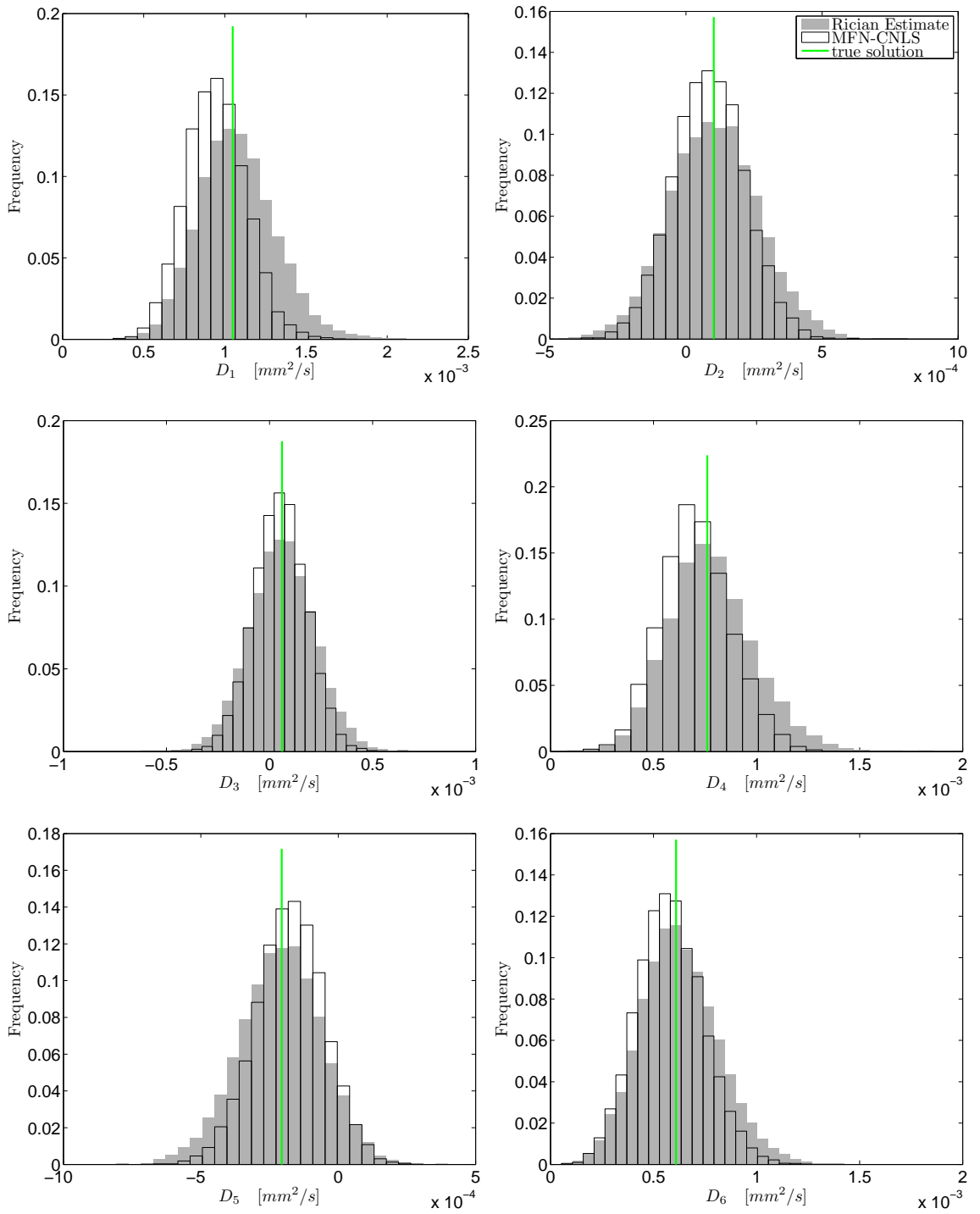


Figure 17: Estimates for the elements of the diffusion tensor.

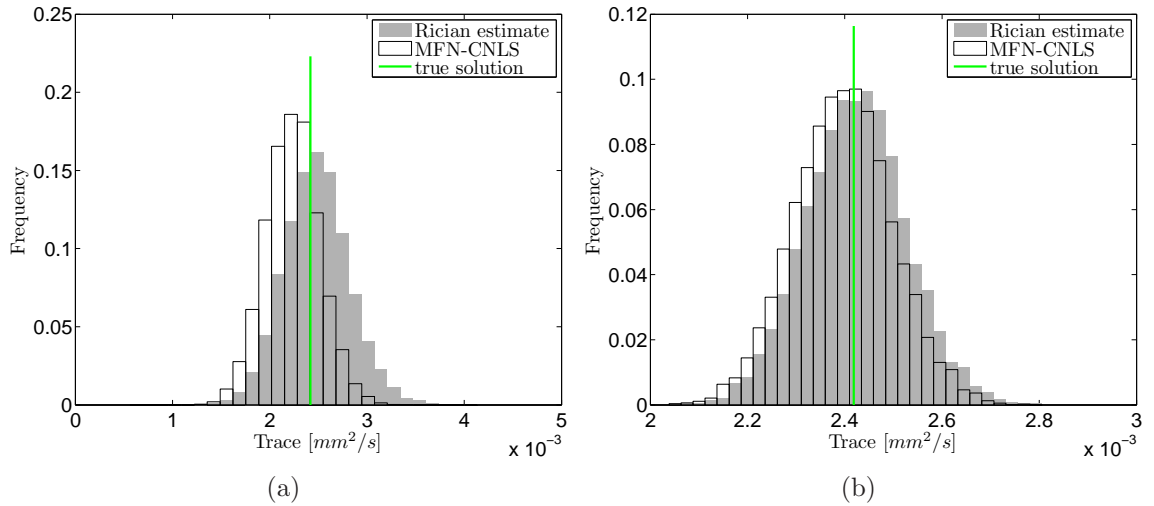


Figure 18: Estimates for the traces of the diffusion tensor for a) SNR = 5 and b) SNR = 15.

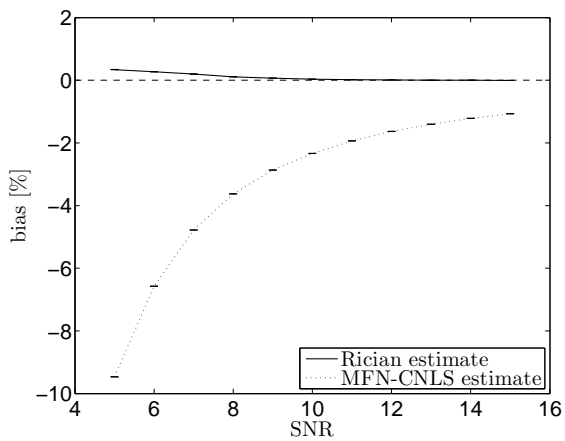


Figure 19: The mean bias for four different tensor traces plotted versus the SNR. It can be seen that the Rician solution is significantly less biased than the MFN-CNLS.

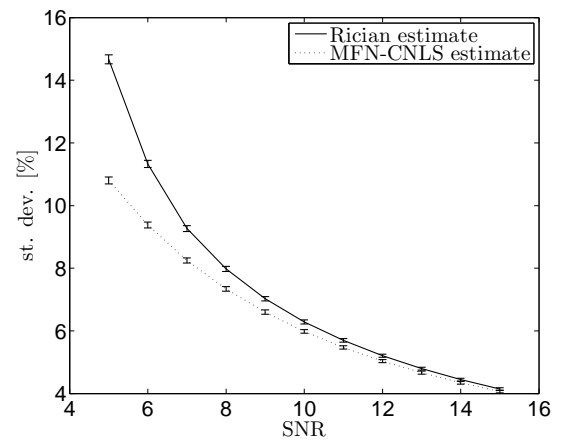


Figure 20: The standard deviation of the error distribution for both methods.

5.3 Error Estimates

The errors for each estimate can be calculated using the inverse of the second derivative matrix as discussed in section 3.5.4. The marginalized error bars are given by the square root of $\nabla^2 L$ while the off-diagonal terms describe the correlation of the parameters.

The reliability of the error estimates was tested by simulated data, each containing 10^4 datasets. Each measurement consists of 69 directed measurements for 23 directions and three different b -factors.

A histogram¹ of the estimated error bars is shown in Fig. 21. Since the data is simulated, the true solution is known and can be used to test the estimated errorbars. The mean of the estimated errorbars was found to be 4.99 %. In the case of Fig. 21 the true solution was 73.61 % of the time within the errorbars. For other simulations with different tensors, this value was always found to be around 70%. This indicates that our error bars are slightly conservative. The fact that this estimation works so well is due to the unimodality of the posterior distribution.

The off-diagonal terms of the covariance matrix describe the correlations between the parameters. It was found that the diagonal terms are negatively correlated, whereas the correlations between off-diagonal/off-diagonal and diagonal/off-diagonal terms are small.

¹ The number of bins for all histograms was calculated by the [optBINS package](#) by Knuth [44]

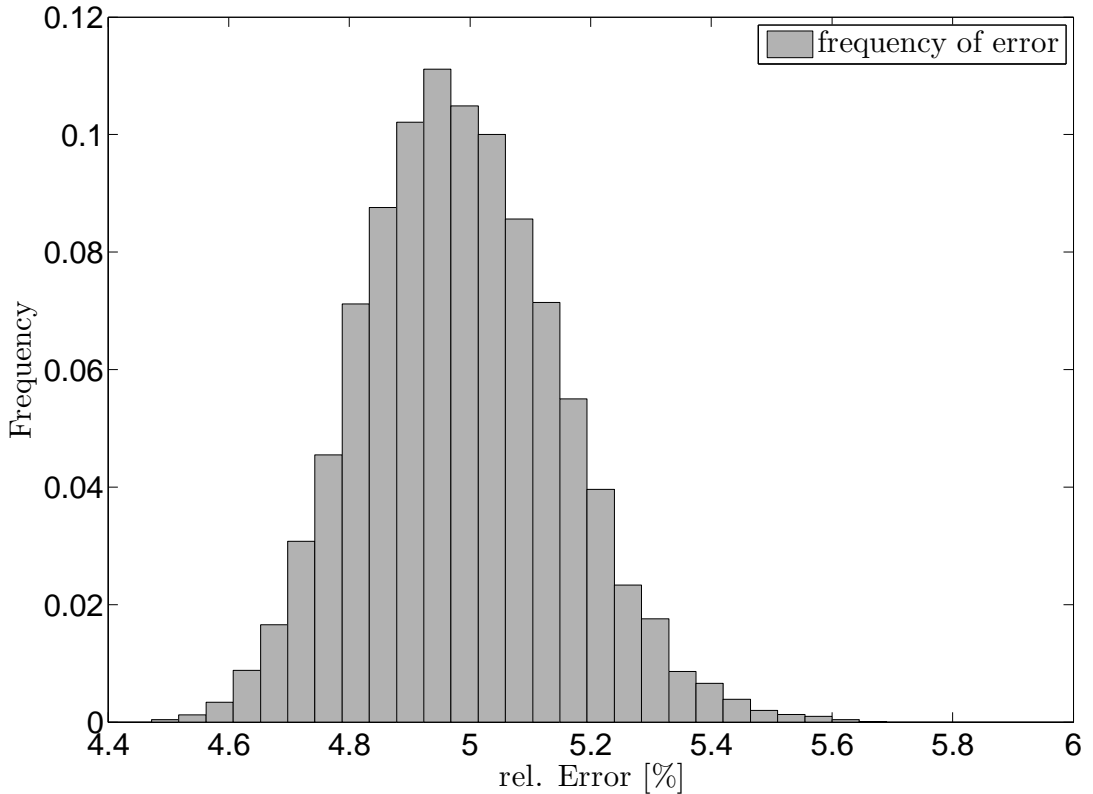


Figure 21: Histogram of the distribution of the error estimates for the covariance method. The mean error is 4.99 %. The true solution lies within the estimated errorbars in 73.61 % of the cases. The SNR of this simulation is 15.

5.4 Conclusions

The method for diffusion tensor estimation presented in this work was shown to be more accurate than the MFN-CNLS method. It does not generally underestimate the diagonal terms since it uses the Rician likelihood function which represents the true distribution of errors in magnitude MRI data. For large signal to noise ratios, both methods produce almost equal results since the Gaussian distribution is the limit of the Rice distribution for large σ . The errorbars were shown to produce reasonable results. In about 70 % of the cases, the true solutions lie within the errorbars, which indicates that our error bar estimates are safely on the conservative side.

One drawback of this method is the reduced speed due to the more complicated

objective function. On a 2 GHz computer, the average time for a SNR= 5 CNLS-MFN estimation was 0.046 s. The Rician method as uncompiled Matlab[®] code takes about 1.119 s for one estimation which is a factor of ≈ 25 slower. The speed and the precision of the estimates might be improved by further optimization of the code. Compiling the code should also yield significant speed improvements.

Bibliography

- [1] P. Lauterbur, Image formation by induced local interactions: examples employing nuclear magnetic resonance, *Nature* 242 (5394) (1973) 190–191.
- [2] R. R. Ernst, W. A. Anderson, Application of fourier transform spectroscopy to magnetic resonance, *Review of Scientific Instruments* 37 (1) (1966) 93–102.
- [3] E. L. Hahn, Spin echoes, *Phys. Rev.* 80 (4) (1950) 580–594.
- [4] P. J. Basser, J. Mattiello, D. LeBihan, Estimation of the effective self-diffusion tensor from the nmr spin echo., *J Magn Reson B* 103 (3) (1994) 247–254.
- [5] H. Haken, H. C. Wolf, *Atom- und Quantenphysik. Einführung in die experimentellen und theoretischen Grundlagen* (Springer Lehrbuch), Springer, Berlin, 2003, english translation available: ISBN 3540208070.
- [6] C. P. Slichter, *Principles of Magnetic Resonance* (Springer Series in Solid-State Sciences), Springer, 1996.
- [7] A. Redfield, On the theory of relaxation processes, *IBM Journal of Research and Development* 1 (1) (1957) 19.
- [8] H. C. Torrey, Bloch equations with diffusion terms, *Phys. Rev.* 104 (3) (1956) 563–565.

- [9] E. O. Stejskal, J. E. Tanner, Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient, *The Journal of Chemical Physics* 42 (1) (1965) 288–292.
- [10] S. Mori, P. B. Barker, Diffusion magnetic resonance imaging: its principle and applications., *Anat Rec* 257 (3) (1999) 102–109.
- [11] P. Kingsley, Introduction to diffusion tensor imaging mathematics: Part i. tensors, rotations, and eigenvectors, *Concepts in Magnetic Resonance Part A* 28 (2) (2006) 101–122.
- [12] P. Kingsley, Introduction to diffusion tensor imaging mathematics: Part ii. anisotropy, diffusion-weighting factors, and gradient encoding schemes, *Concepts in Magnetic Resonance Part A* 28 (2) (2006) 123–154.
- [13] P. Kingsley, Introduction to diffusion tensor imaging mathematics: Part iii. tensor calculation, noise, simulations, and optimization, *Concepts in Magnetic Resonance Part A* 28 (2) (2006) 155–179.
- [14] A. Fick, Über Diffusion, *Annalen der Physik* 170 (1855) 59–86.
- [15] A. Einstein, Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen, *Annalen der Physik* 17 (4) (1905) 548–560.
- [16] E. T. Jaynes, Clearing up mysteries - the original goal, in: J. Skilling (Ed.), *Maximum Entropy and Bayesian Methods*, Vol. 8, 1989, pp. 1–27.
- [17] D. L. Thomas, M. F. Lythgoe, G. S. Pell, F. Calamante, R. J. Ordidge, The measurement of diffusion and perfusion in biological systems using magnetic resonance imaging, *Physics in Medicine and Biology* 45 (8) (2000) R97–R138.
- [18] P. J. Basser, J. Mattiello, D. LeBihan, MR diffusion tensor spectroscopy and imaging., *Biophys J* 66 (1) (1994) 259–267.

- [19] G. L. Bretthorst, C. D. Kroenke, J. J. Neil, Characterizing water diffusion in fixed baboon brain, *Bayesian Inference and Maximum Entropy Methods In Science And Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering* 735 (1) (2004) 3–15.
- [20] D. L. Bihan, J. F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, H. Chabriat, Diffusion tensor imaging: concepts and applications., *J Magn Reson Imaging* 13 (4) (2001) 534–546.
- [21] R. Bhatia, *Positive Definite Matrices (Princeton Series in Applied Mathematics)*, Princeton University Press, 2006.
- [22] D. L. Bihan, P. van Zijl, From the diffusion coefficient to the diffusion tensor., *NMR Biomed* 15 (7-8) (2002) 431–434.
- [23] M. E. Moseley, Y. Cohen, J. Mintorovitch, L. Chilewitt, H. Shimizu, J. Kucharczyk, M. F. Wendland, P. R. Weinstein, Early detection of regional cerebral ischemia in cats: comparison of diffusion- and t2-weighted mri and spectroscopy., *Magn Reson Med* 14 (2) (1990) 330–346.
- [24] M. Kubicki, C.-F. Westin, S. E. Maier, H. Mamata, M. Frumin, H. Ersner-Hershfield, R. Kikinis, F. A. Jolesz, R. McCarley, M. E. Shenton, Diffusion tensor imaging and its application to neuropsychiatric disorders., *Harv Rev Psychiatry* 10 (6) (2002) 324–336.
- [25] K. Arfanakis, M. Gui, M. Lazar, Optimization of white matter tractography for pre-surgical planning and image-guided surgery., *Oncol Rep* 15 Spec no. (2006) 1061–1064.
- [26] B. A. Ardekani, J. Nierenberg, M. J. Hoptman, D. C. Javitt, K. O. Lim, MRI study of white matter diffusion anisotropy in schizophrenia., *Neuroreport* 14 (16) (2003) 2025–2029.

- [27] H. Gudbjartsson, S. Patz, The Rician distribution of noisy MRI data., *Magn Reson Med* 34 (6) (1995) 910–914.
- [28] R. M. Henkelman, Measurement of signal intensities in the presence of noise in mr images., *Med Phys* 12 (2) (1985) 232–233.
- [29] M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, Dover Publications, 1965.
- [30] S. Rice, Mathematical analysis of random noise, *Selected Papers on Noise and Stochastic Processes* 24 (1954) 133–294.
- [31] D. Sivia, J. Skilling, *Data Analysis: A Bayesian Tutorial*, Oxford University Press, USA, 2006.
- [32] W. Gosset, The probable error of a mean, *Biometrika* 6 (1) (1908) 1–25.
- [33] T. Bayes, An essay towards solving a problem in the doctrine of chances, *Phil. Trans. Roy. Soc* 53 (1764) 370–418.
- [34] E. T. Jaynes, G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003.
- [35] A. Caticha, A. Giffin, Updating probabilities, in: *Bayesian Inference and Maximum Entropy Methods In Science and Engineering*, Vol. 872 of American Institute of Physics Conference Series, 2006, pp. 31–42.
- [36] 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., *J Magn Reson* 182 (1) (2006) 115–125.
- [37] C. G. Koay, J. D. Carew, A. L. Alexander, P. J. Basser, M. E. Meyerand, Investigation of anomalous estimates of tensor-derived quantities in diffusion tensor imaging., *Magn Reson Med* 55 (4) (2006) 930–936.

- [38] I. N. Bronstein, K. A. Semendjajew, G. Musiol, Taschenbuch der Mathematik, Deutsch (Harri), 2005, english translation available: ISBN 0442211716.
- [39] G. E. Shilov, Linear Algebra, Dover Publications, 1977.
- [40] S. A. Schelkunoff, A note on a paper concerning a bessel function, The American Mathematical Monthly 37 (9) (1930) 491–492.
- [41] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, 1992.
- [42] M. T. Heath, Scientific Computing: An Introductory Survey, McGraw-Hill Companies, 1996.
- [43] D. K. Jones, M. A. Horsfield, A. Simmons, Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging., Magn Reson Med 42 (3) (1999) 515–525.
- [44] K. Knuth, Optimal data-based binning for histograms (2006).

Appendix A

Marginalized Rice Distribution

If the standard deviation of a Gaussian probability distribution is unknown, it can be marginalized with Jeffrey's Prior resulting in the Student-t distribution. This section shows the marginalization of the Rice distribution over σ . The marginalization integral with Jeffreys Prior (given by $1/\sigma$) is

$$P(M|A) = \int_0^\infty \frac{1}{\sigma} \frac{M}{\sigma^2} e^{-\frac{M^2+A^2}{2\sigma^2}} I_0\left(\frac{A M}{\sigma^2}\right) d\sigma \quad (\text{A.1})$$

$\frac{M A}{\sigma^2}$ is substituted by x :

$$dx = -\frac{2 M \cdot A}{\sigma^2} d\sigma = -\frac{2}{\sqrt{M \cdot A}} x^{\frac{3}{2}} d\sigma \quad (\text{A.2})$$

The minus sign is absorbed by changing the direction of integration.

$$P(M|A) = \frac{1}{2A} \int_0^\infty e^{-\frac{M^2+A^2}{2\sigma^2} x} \cdot I_0\left(\frac{M A}{\sigma^2}\right) dx \quad (\text{A.3})$$

For simplicity, let $\alpha = \frac{M^2+A^2}{\sigma^2}$.

According to [29], the modified Bessel function of the first kind can be written as

$$I_0(x) = \frac{1}{\pi} \int_0^\pi e^{x \cdot \cos \theta} d\theta. \quad (\text{A.4})$$

Substituting into Eq. (A.3) yields

$$P(M|A) = \frac{1}{2\pi A} \int_0^\infty \int_0^\pi e^{-\alpha x} e^{x \cos \theta} dx d\theta. \quad (\text{A.5})$$

Now the integration over x can be carried out:

$$\begin{aligned} P(M|A) &= \frac{1}{2\pi A} \int_0^\pi \left\{ -\frac{1}{\alpha - \cos \theta} \cdot [e^{-x(\alpha - \cos \theta)}]_0^\infty \right\} d\theta \\ &= \frac{1}{2\pi A} \int_0^\pi \frac{1}{\alpha - \cos \theta} d\theta \end{aligned} \quad (\text{A.6})$$

This integral can be found in [38] which leads to

$$P(M|A) = \frac{1}{2\pi A} \cdot \left[\frac{2 \arctan \left(\frac{(1+\alpha) \tan \frac{\theta}{2}}{\sqrt{\alpha^2 - 1}} \right)}{\sqrt{\alpha^2 - 1}} \right]_0^\pi. \quad (\text{A.7})$$

Further simplification leads to the simple final result:

$$P(M|A) = \frac{M}{|M^2 - A^2|} \quad (\text{A.8})$$

However this distribution is not normalizable since the integral over M does not converge.

Appendix B

Table of Acronyms

CNLS	Constrained Nonlinear Least Squares
DTI	Diffusion Tensor Imaging
DWI	Diffusion Weighted Imaging
fMRI	functional Magnetic Resonance Imaging
MFN	Modified Full Newton
MRI	Magnetic Resonance Imaging
NMR	Nuclear Magnetic Resonance
RF	Radio Frequency
SNR	Signal to Noise Ratio